BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

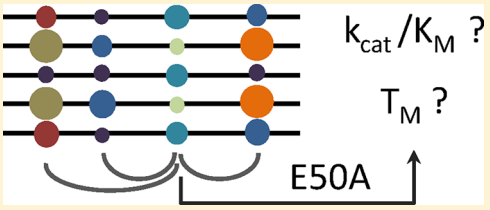
Experimental Assessment of the Importance of Amino Acid Positions Identified by an Entropy-Based Correlation Analysis of Multiple-Sequence Alignments

Susanne Dietrich, Nadine Borst, Sandra Schlee, Daniel Schneider, Jan-Oliver Janda, Reinhard Sterner,* and Rainer Merkl*

Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany

Supporting Information

ABSTRACT: The analysis of a multiple-sequence alignment (MSA) with correlation methods identifies pairs of residue positions whose occupation with amino acids changes in a concerted manner. It is plausible to assume that positions that are part of many such correlation pairs are important for protein function or stability. We have used the algorithm H2r to identify positions k in the MSAs of the enzymes anthranilate phosphoribosyl transferase (AnPRT) and indole-3-glycerol phosphate synthase (IGPS) that show a high conn(k) value, i.e., a large number of significant correlations in which k is involved. The importance of the identified residues was experimentally validated by performing mutagenesis studies with sAnPRT and sIGPS from the archaeon *Sulfolobus solfataricus*. For sAnPRT, five H2r mutant proteins were generated by replacing nonconserved residues with alanine or the prevalent residue of the MSA. As a control, five residues with conn(k) values of zero were chosen randomly and replaced with alanine. The catalytic activities and conformational stabilities of the H2r and control mutant proteins were analyzed by steady-state enzyme kinetics and thermal unfolding studies. Compared to wild-type sAnPRT, the catalytic efficiencies (k_{cat}/K_M) were largely unaltered. In contrast, the apparent thermal unfolding temperature (T_M^{app}) was lowered in most proteins. Remarkably, the strongest observed destabilization (ΔT_M^{app} = 14 °C) was caused by the V284A exchange, which pertains to the position with the highest correlation signal [conn(k) = 11]. For sIGPS, six H2r mutant and four control proteins with alanine exchanges were generated and characterized. The k_{cat}/K_M values of four H2r mutant proteins were reduced between 13- and 120-fold, and their T_M^{app} values were decreased by up to 5 °C. For the sIGPS control proteins, the observed activity and stability decreases were much less severe. Our findings demonstrate that positions with high conn(k) values have an increased probability of being important for enzyme function or stability.



The analysis of a multiple-sequence alignment (MSA) of homologous proteins provides information about the significance of individual residues. Strictly conserved residues are in most cases essential for protein function,[1,2] whereas a prevalent but not exclusively found amino acid is often important for protein stability.[3,4] Moreover, it has been realized that the analysis of the mutual interdependence of residue frequencies found in several columns of an MSA can also be instructive.[5,6] For instance, the occurrence of a specific amino acid at a given position of the MSA may crucially depend on the local environment, which can impose restrictions with respect to the size or chemical properties of the neighboring residues. Thus, an amino acid replacement at one position is tolerated only together with a complementary residue substitution at a correlated site. As a consequence, the frequencies of particular residues at adjacent positions in the structure of a protein can be interrelated. This coupling can be detected by a column-wise correlation analysis of the MSA based on information theory; classical approaches have been reviewed.[7] In addition to direct contacts, a correlation of residue frequencies can also be due to the concerted action of amino acids that are far from each other in the three-dimensional structure. For example, it has been shown that coupled motions of remote residues can be crucial for enzyme catalysis.[8,9] Recent in silico analyses have demonstrated that correlated mutations tend to be over-represented near active site residues[10,11] and that point mutations at coevolving sites are associated with human diseases more frequently than expected.[12]

An important practical application for correlation analysis is protein design. Complementary to the consensus approach, which assumes that the most frequent residue observed in the respective column of an MSA is the most stabilizing one, the detection of covariation can contribute to the identification of residues that are important for the integrity of a protein.[13] For example, it has been demonstrated that an energy function, which encompasses coevolution between amino acid residues, is able to specify sequences that fold into a compact three-dimensional structure.[14]

The majority of the established protocols for the identification of correlated mutations are based on the computation of a global covariation signal.[5,9,15−19] Related to

Received: June 6, 2012
Revised: June 26, 2012
Published: June 27, 2012

this concept, "perturbation" methods have been introduced.[6] In the first step of a perturbation analysis, for a given position k, the amino acid aa_i^k occurring most frequently at position k is identified. Then, a correlation signal between k and any other position l is calculated. To this end, the mean residue composition at l is compared with the residue composition at l for the sequences possessing aa_i^k.[20] In combination with molecular dynamics simulations, perturbation analysis has been successfully applied to predict residues essential for catalysis that are coupled by coevolution and motion.[9] However, perturbation analysis has been under debate[21] because it ignores all information contained in sequences not possessing aa_i^k. In addition, it is unclear whether this approach assesses in an appropriate manner the strong phylogenetic signal that interferes with the signal originating from structural or functional coupling.[22]

In an attempt to advance correlation methods, we have recently implemented the algorithm H2r that is based on Shannon's information theory.[23] As schematically outlined in Figure 1, H2r uses the information content of the complete set

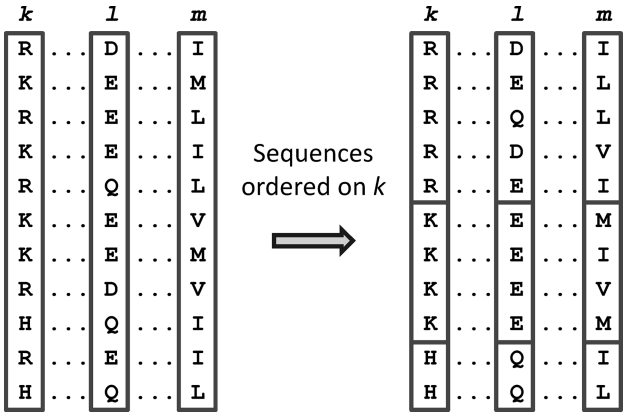


Figure 1. Illustration of correlation analysis as implemented in H2r. The residues found in three columns of an MSA representing positions k, l, and m are shown in the left panel. In the right panel, the sequences are sorted according to the composition of column k to highlight mutual interdependencies. The frequencies $f(aa_i^k,aa_i^l)$ of amino acid pairs aa_i and aa_j observed in columns k and l are biased, as only a small number of specific combinations occur. Using these frequencies, the term $U(k,l)$ is computed, which is a normalized measure for the correlation between the content of columns k and l (for details, see ref 23). This measure is a symmetric one; hence, $U(k,l) = U(l,k)$. No correlation is detectable for the k, m pair. A perturbation method would compare only the mean residue frequencies at column l with the frequencies observed in those sequences that carry in column k the most frequent amino acid (here R). Thus, correlation signals related to less abundant residues would be ignored. Note that without removing highly similar sequences, an overrepresentation of amino acid pairs could be due to a close phylogenetic relationship and might cause false positive predictions. A prerequisite for the computation of $U(k,l)$ values is a sufficiently large MSA. To deduce the pairwise frequencies $f(aa_i^k,aa_i^l)$ with acceptable precision, the MSA has to consist of at least 125 sequences.[18]

of sequences of an MSA to identify correlations between amino acid frequencies. The algorithm computes the term $U(k,l)$, which is a normalized measure for mutual information and quantifies the degree of coupling between positions k and l of an MSA. Using a large set of synthetic MSAs, we have shown that $U(k,l)$ values identify correlated positions in a manner independent of a specific pattern of residue conservation.[23] However, pairs of positions with strictly conserved residues are excluded from the analysis because they per se do not contain a correlation signal. Because we have noted that in alignments of native sequences those positions with high $U(k,l)$ values are interlinked, we utilize concepts of network analysis for their characterization. Specifically, all $U(k,l)$ values derived from an MSA are rank-ordered, and the N highest $U_N(k,l)$ values are used to determine the connectivity conn(k) for each position k. We define conn(k) as the number of high-scoring pairs $U_N(k,l)$ of which position k is an element; the value of N is chosen according to the size of the protein. Figure 2 illustrates how conn(k) values are calculated.

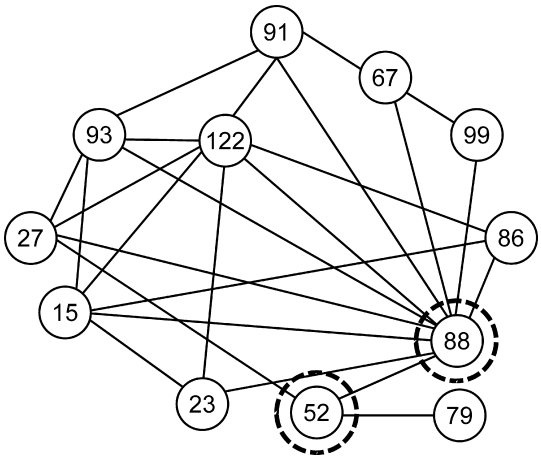


Figure 2. Scheme illustrating the determination of conn(k) values. For the shown virtual example, 25 $U(k,l)$ values were used to create the graph. Each node represents a residue k (or l), and each vertex is due to a $U_{25}(k,l)$ value. Nodes differ by their conn(k) value, which is the number of vertices ending in k. Here, $conn_{25}$ of residue 52 [$conn_{25}(52)$] is 3 and $conn_{25}(88)$ is 10. Note that H2r uses at least the 75 most extreme $U(k,l)$ values to determine conn(k).

By computing conn(k), we achieve two objectives. (i) As we focus on those positions k that cause the strongest correlation signals, we reduce the risk of predicting false positives. Our simulations have shown that conn(k) values of >3 are highly unlikely when analyzing an MSA of random composition with default settings of H2r. (ii) Rather than merely focusing on the coupling between k, l pairs, we are also able to assess individual positions k with respect to their significance for the protein.[23] A recent independent report confirmed that conn(k) is a robust indicator for the detection of covariation signals between residue positions.[24] H2r is accessible as a web service at http://www-bioinf.uni-regensburg.de.

In this work, we describe the first experimental validation of predictions made by H2r. For this purpose, we have performed a mutagenesis study with two well-characterized enzymes from the tryptophan biosynthetic pathway, anthranilate phosphoribosyl transferase (sAnPRT; encoded by the *strp*D gene) and indole-3-glycerol phosphate synthase (sIGPS; encoded by the *strp*C gene) from the hyperthermophilic archaeon *Sulfolobus solfataricus*. Both sAnPRT and sIGPS are highly thermostable but only marginally active at room temperature.[25,26] Crystal structure analysis has shown that sAnPRT is a homodimer with a subunit fold similar to that of nucleoside phosphorylases, and that sIGPS is a monomer with the ubiquitous $(\beta\alpha)_8$-barrel enzyme fold.[27,28] Previous site-directed mutagenesis experiments have led to the identification of amino acids crucial for

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

dimer formation of sAnPRT and of conserved residues essential for substrate binding and catalysis of sAnPRT and sIGPS.[29−31] We have now analyzed the significance of nine positions of sAnPRT and sIGPS, which were assigned conn(k) values of >3 by H2r but are not conserved among the known AnPRT and IGPS proteins. The amino acids at these positions were replaced with either alanine or the prevalent residue found in the respective column of the MSA. As a control, nine positions of sAnPRT and sIGPS with a conn(k) value of zero were chosen and replaced with alanine. The recombinant sAnPRT and sIGPS mutant proteins were produced in *Escherichia coli*, purified, and analyzed by thermal unfolding and steady-state enzyme kinetics. The results show that the effects leading to a reduced thermal stability in the case of sAnPRT or a diminished catalytic activity in the case of sIGPS are much stronger at positions with high conn(k) values than at positions with a conn(k) value of zero. These findings confirm that H2r identifies important nonconserved residue positions of a protein without requiring structural information.

EXPERIMENTAL PROCEDURES

Computational Methods. Two MSAs were generated for AnPRT and IGPS proteins by using BLAST[32] to extract sequences from the NCBI database[33] and by aligning the output with MAFFT.[34] H2r was utilized with default cutoffs of 20 and 90% sequence identity to eliminate too dissimilar and too similar sequences, respectively.[23] The program determined $U(k,l)$ values for all columns k and l of an MSA that were <90% conserved. $U(k,l)$ values indicate the coupling in the pairwise amino acid distribution of residue positions k and l. The higher the $U(k,l)$, the stronger the mutual dependency of the residue composition. On the basis of the 75 most extreme $U(k,l)$ values, conn(k) values were determined for all residues of a protein. Residue positions with a conn(k) value of >3 were considered for H2r mutant proteins and those with a conn(k) value of zero for control mutants.

Cloning, Expression, and Purification of Recombinant Proteins. To introduce amino acid exchanges into sAnPRT and sIGPS, *strp*D and *strp*C mutant genes were generated by mega-primer[35] or overlap-extension[36] polymerase chain reaction using the oligonucleotides listed in Table S1 of the Supporting Information, together with flanking oligonucleotides specific for the used expression plasmids. The *strp*D mutants were cloned into the pQE40 plasmid (Quiagen) using the restriction sites for *Bam*HI and *Hin*dIII, and the *strp*C mutants were cloned into the pET21a(+) plasmid (Stratagene) using the restriction sites for *Nhe*I and *Xho*I. As a consequence, the recombinant sAnPRT and sIGPS mutant proteins carried N-terminal and C-terminal hexahistidine tags, respectively. The *strp*D mutants were expressed heterologously in *E. coli* strain W3113 *trp*EA, which contained the repressor plasmid pDMI,1,[37] and the *strp*C mutants were expressed in *E. coli* strain BL21-CodonPlus(DE3) RIPL (Stratagene). The recombinant sAnPRT and sIGPS mutant and wild-type proteins were purified by heat precipitation of the *E. coli* host proteins and metal chelate affinity chromatography.[29,38] Additionally, sAnPRT proteins were purified by ion exchange chromatography using a Resource Q column (1 mL, GE Healthcare). Bound protein was eluted from the pre-equilibrated column with a linear gradient from 0 to 750 mM KCl in 50 mM Tris-HCl (pH 9.0). Fractions containing pure sAnPRT and sIGPS were identified by sodium dodecyl sulfate−polyacrylamide gel electrophoresis, pooled, and dialyzed. The yields per liter of cell suspension were between 0.05 and 0.3 mg for the sAnPRT mutant proteins and between 5 and 30 mg for the sIGPS mutant proteins. The proteins were dropped into liquid nitrogen and stored at −80 °C. Protein concentrations were determined by measuring the absorbance at 280 nm, using molar extinction coefficients that were calculated from the amino acid sequence.[39]

Protein Characterization. The AnPRT reaction was monitored under steady-state conditions at 37 °C by a fluorimetric assay performed in 50 mM Tris-HCl (pH 7.5)[40] in the presence of 200 μM $MgCl_2$. The IGPS reaction was monitored under steady-state conditions at 25 °C by a fluorimetric assay performed in 50 mM HEPPS-KOH (pH 7.5) containing 4 mM EDTA.[38] The Michaelis constants (K_M^{AA} and K_M^{PRPP} for sAnPRT and K_M^{CdRP} for sIGPS) were determined by fitting a hyperbolic equation (using SigmaPlot, Systat Software Inc.) to saturation curves of the various substrates. For sAnPRT, saturation curves for anthranilate (AA) and PRPP were recorded in the presence of an excess ($10K_M$) of the second substrate. The turnover numbers (k_{cat}) were determined by dividing the fitted maximum velocity (V_{max}) by the concentration of active sites ($[E_0]$).

Thermal denaturation of sAnPRT and sIGPS dissolved in 50 and 10 mM potassium phosphate (pH 7.5), respectively, was monitored with a JASCO J-815 circular dichroism (CD) spectrometer (Jasco, Easton, MD) in a 0.1 cm cuvette by following the loss of ellipticity at 220 nm. Unfolding was induced by increasing the temperature in 0.1 °C increments at a ramp rate of 1 °C/min using a Peltier-effect temperature controller. The measured ellipticity values were normalized, and the midpoint temperature of the unfolding transition (T_M^{app}) was determined.

The association states of sAnPRT and sIGPS mutant proteins were determined by analytical gel filtration chromatography using a Superdex 75 column. Elution was performed at 25 °C with a flow rate of 0.5 mL/min on a calibrated Superdex 75 column (1 cm × 30 cm, GE Healthcare) that was equilibrated with 50 mM potassium phosphate (pH 7.5) containing 300 mM KCl.

RESULTS

Identification by H2r of Putatively Important Residue Positions in sAnPRT and sIGPS. For the identification of important residue positions in sAnPRT, 412 AnPRT sequences were collected with BLAST[32] and aligned in an MSA. To eliminate dissimilar sequences that might introduce noise as well as highly similar sequences that might cause a correlation signal because of their close phylogenetic relationship, pairwise comparisons were performed with sAnPRT as the first reference. A sequence was removed if the comparison yielded <20 or >90% identical residues. An MSA composed of the 288 remaining AnPRTs was analyzed with H2r, which identified five positions with conn(k) values of >3 (Table S2 of the Supporting Information). The corresponding residues of sAnPRT, glutamate 50, glutamate 54, isoleucine 235, valine 284, and phenylalanine 297, were replaced with alanine (Table 1). Alanine is the standard exchange in mutagenesis studies as it is chemically inert and does not introduce charge or steric constraints. Additionally, in cases where a clear amino acid preference deviating from wild-type sAnPRT could be deduced from the MSA, the most frequently encountered ("consensus") residue was also introduced because it has been shown that this strategy can lead to increased protein stability.[3,4] This

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

Table 1. Experimentally Validated Positions of sAnPRT and sIGPS[a]

protein	residue at identified position	conn(k)	most frequent residue	occurrence (%)	conservation rank[b]
sAnPRT	E 50	4	E	77.5	62
	E 54	4	E	90.5	33
	I 235	4	E	50.7	149
	V 284	11	A	19.7	309
	F 297	4	Y	38.4	211
	E 123	0	E	69.1	82
	K 203	0	E	33.8	234
	F 251	0	F	41.7	194
	I 316	0	I	60.3	117
	K 318	0	S	65.1	100
sIGPS	W 8	10	K	73.7	65
	R 54	7	K	64.1	88
	R 64	14	F	62.0	89
	S 181	4	N	86.2	44
	S 56	0	S	99.5	9
	S 58	0	S	92.6	34
	L 187	0	F	78.8	54
	V 208	0	V	73.3	67

[a]For sAnPRT (331 residues), five wild-type residues with the highest conn(k) values were replaced by site-directed mutagenesis, both with alanine and with the most frequent (consensus) residue of the MSA. These residues constitute the H2r mutant proteins. As a control, five randomly selected residues with a conn(k) value of zero were also replaced with alanine. For sIGPS (248 residues), from the 13 positions with conn(k) values of >3 (cf. Table S2 of the Supporting Information), two wild-type residues located in the proximity of the substrate CdRP within the crystal structure (PDB entry 1LBF) were exchanged. In addition, residues at positions 64 and 54 were replaced together on the basis of their high conn(k) and U(54,64) values. As a control, four residues in the proximity of CdRP with a conn(k) value of zero were selected. [b]The value is the rank of the positions according to their conservation in the MSAs.

consideration led to the replacement of isoleucine 235 with glutamate and phenylalanine 297 with tyrosine. At positions 50 and 54, no additional exchanges were introduced because at both sites sAnPRT contained the consensus residue glutamate. At position 284, the prevalent residue was alanine, albeit with a frequency of only 19.7% (Table 1). In the following, we name a set of proteins with amino acid exchanges selected on the basis of a conn(k) value of >3 "H2r mutant proteins". As a control, five positions with a conn(k) value of zero and a degree of conservation comparable to that observed in the H2r mutant proteins were chosen stochastically from all 331 positions with the help of a random number generator. The corresponding residues of sAnPRT, glutamate 123, lysine 203, phenylalanine 251, isoleucine 316, and lysine 318, were replaced with alanine, and the resulting proteins served as a reference for the effects observed with the H2r mutant proteins.

For the identification of potentially important residues of sIGPS, we generated an MSA composed of 567 IGPS sequences. Its processing by H2r identified 13 positions with a conn(k) value of >3 (Table S3 of the Supporting Information). Four of them were tested by site-directed mutagenesis based on the following rationale: Previously, a number of putative active site residues, which were selected because of their conservation, were analyzed with respect to their role in catalysis using IGPS from *E. coli*.[41] We now tested also nonconserved residues, which were identified by H2r and are located close to the active site of sIGPS. On the basis of this criterion, we replaced tryptophan 8 and serine 181 with alanine. In addition, tryptophan 8 was replaced with lysine and serine 181 with asparagine, which are the consensus residues at these two positions of the MSA (Table 1). As a control for serine 181, serines 56 and 58 were selected, and as a control for tryptophan 8, the hydrophobic residues leucine 187 and valine 208 were chosen and replaced with alanine. These residue positions have conn(k) values of zero, are at least as conserved as the residue positions selected for the two H2r mutant proteins, and are located in similar proximity of the active site.

Table 2. Catalytic Activities and Thermal Stabilities of sAnPRT Mutant Proteins[a]

	sAnPRT protein	k_{cat} (s^{-1})	K_M^{PRPP} (μM)	K_M^{AA} (nM)	k_{cat}/K_M^{PRPP} (M^{-1} s^{-1})	k_{cat}/K_M^{AA} (M^{-1} s^{-1})	effect on activity[b]	T_M^{app} (°C)	ΔT (°C)
H2r mutant proteins	WT	0.35 ± 0.013	127 ± 23	38 ± 2	2.7×10^3	9.2×10^6	–	91	–
	E50A	0.37 ± 0.017	278 ± 52	70 ± 8	1.3×10^3	5.3×10^6	≤2, ≤2	87	−4
	E54A	0.44 ± 0.025	203 ± 43	96 ± 10	2.2×10^3	4.6×10^6	≤2, ≤2	86	−5
	I235A	0.32 ± 0.008	146 ± 18	24 ± 1	2.2×10^3	13.3×10^6	≤2, 0.7	85	−6
	I235E	0.38 ± 0.019	367 ± 65	44 ± 3	1.0×10^3	8.6×10^6	3, ≤2	80	−11
	V284A	0.24 ± 0.006	180 ± 19	53 ± 4	1.3×10^3	4.5×10^6	≤2, ≤2	77	−14
	F297A	0.32 ± 0.015	251 ± 46	12 ± 2	1.3×10^3	26.7×10^6	≤2, 0.3	91	0
	F297Y	0.28 ± 0.009	373 ± 49	31 ± 4	0.8×10^3	9.0×10^6	3, ≤2	91	0
controls	WT	0.12 ± 0.006	163 ± 32	9 ± 3	7.4×10^2	1.3×10^7	–	91	–
	E123A	0.11 ± 0.003	57 ± 9	33 ± 8	1.9×10^3	3.4×10^6	4, 4	83	−8
	K203A	0.05 ± 0.004	560 ± 140	18 ± 3	0.9×10^2	2.8×10^6	8, 5	90	−1
	F251A	0.04 ± 0.002	147 ± 42	24 ± 4	2.7×10^2	1.7×10^6	3, 8	82	−9
	I316A	0.03 ± 0.001	116 ± 20	9 ± 1	2.6×10^2	3.4×10^6	3, 4	89	−2
	K318A	0.05 ± 0.003	530 ± 75	9 ± 1	0.9×10^2	5.6×10^6	8, 2	90	−1

[a]The steady-state constants k_{cat} and K_M were determined by the analysis of substrate saturation curves, which were constructed from initial velocities measured at different concentrations of anthranilate (AA) and PRPP, in the presence of saturating concentrations of the second substrate (>10K_M). Experiments were conducted in 50 mM Tris-HCl (pH 7.5) at 37 °C in the presence of 200 μM $MgCl_2$. The shown mean values and standard deviations were deduced from the analysis of multiple measurements. Steady-state constants for wild-type sAnPRT (WT) were determined with two different protein preparations. The apparent melting temperatures (T_M^{app}) were determined by the analysis of thermal unfolding traces monitored by far-UV CD spectroscopy. The accuracy of the given values is ±1 °C. Experiments were conducted in 50 mM potassium phosphate (pH 7.5). [b]The values are the $(k_{cat}/K_M^{PRPP})_{wild\text{-}type}/(k_{cat}/K_M^{PRPP})_{mutant}$ and $(k_{cat}/K_M^{AA})_{wild\text{-}type}/(k_{cat}/K_M^{AA})_{mutant}$ ratios.

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

Table 3. Catalytic Activities and Thermal Stabilities of sIGPS Mutant Proteins[a]

	sIGPS protein	k_{cat} (s^{-1})	K_M^{CdRP} (μM)	k_{cat}/K_M^{CdRP} (M^{-1} s^{-1})	effect on activity[b]	T_M^{app} (°C)	ΔT (°C)
H2r mutant proteins	WT	0.06 ± 0.002	0.05 ± 0.006	1.20×10^6	–	92	–
	W8A	0.13 ± 0.002	1.40 ± 0.08	0.09×10^6	13	90	−2
	W8K	0.31 ± 0.01	0.37 ± 0.03	0.84×10^6	≤2	87	−5
	S181A	0.008 ± 0.003	0.59 ± 0.09	0.01×10^6	120	93	1
	S181N	0.39 ± 0.03	0.69 ± 0.11	0.56×10^6	≤2	90	−2
	R54A/R64A	0.17 ± 0.008	2.70 ± 0.40	0.06×10^6	20	94	2
	R54K/R64F	0.14 ± 0.003	1.80 ± 0.13	0.08×10^6	15	94	2
controls	S56A	0.13 ± 0.009	0.16 ± 0.03	0.81×10^6	≤2	91	−1
	S58A	0.09 ± 0.004	0.17 ± 0.02	0.53×10^6	≤2	93	1
	L187A	0.13 ± 0.007	0.19 ± 0.04	0.68×10^6	≤2	91	−1
	V208A	0.21 ± 0.007	0.21 ± 0.02	1.00×10^6	0.8	88	−4

[a]The steady-state constants k_{cat} and K_M^{CdRP} were determined by the analysis of substrate saturation curves, which have been constructed from initial velocities measured at different concentrations of CdRP. Experiments were conducted in 50 mM HEPPS-KOH (pH 7.5) at 25 °C in the presence of 4 mM EDTA. The shown mean values and standard deviations were deduced from the analysis of multiple measurements. The apparent melting temperatures (T_M^{app}) were determined by the analysis of thermal unfolding traces monitored by far-UV CD spectroscopy. The accuracy of the given values is ±1 °C. Experiments were conducted in 10 mM potassium phosphate (pH 7.5). [b]The value is the $(k_{cat}/K_M^{CdRP})_{wild\text{-}type}/(k_{cat}/K_M^{CdRP})_{mutant}$ ratio.

Remarkably, positions 54 and 64 are characterized by large conn(k) values of 7 and 14, respectively, and a high U(54,64) value (not shown), which indicates a strong correlation. Therefore, we also generated proteins that carry at these positions either alanines (R54A/R64A) or the consensus residues (R54K/R64F) (Table 1).

Catalytic Activity and Thermal Stability of sAnPRT and sIGPS Mutant Proteins. Nucleotide exchanges leading to the planned amino acid substitutions in sAnPRT and sIGPS were introduced into the *strp*D and *strp*C genes by site-directed mutagenesis, and the plasmid-encoded genes were expressed in *E. coli*. The recombinant proteins were purified by heat precipitation of the host proteins, followed by metal-chelate affinity chromatography using a hexahistidine tag, and an additional ion exchange chromatographic step in the case of sAnPRT. Analytical gel filtration demonstrated that all sAnPRT mutant proteins form homogeneous homodimers and that all sIGPS mutant proteins form homogeneous monomers.

The catalytic activities of the sAnPRT and sIGPS H2r mutant and control proteins were analyzed by steady-state enzyme kinetics. AnPRT catalyzes the Mg^{2+}-dependent conversion of anthranilate and PRPP into *N*-(5′-phosphoribosyl)anthranilate and pyrophosphate. To determine the turnover numbers (k_{cat}) and the Michaelis constants for anthranilate (K_M^{AA}) and PRPP (K_M^{PRPP}), saturation curves were determined for both substrates. Table 2 shows that the steady-state constants of the H2r mutant proteins are similar to those of wild-type sAnPRT, which also holds for the control proteins carrying the E123A and I316A exchanges. For control proteins K203A and K318A, k_{cat}/K_M^{PRPP} is reduced 8-fold, and for control protein F251A, k_{cat}/K_M^{AA} is decreased by the same factor.

IGPS catalyzes the conversion of 1-(*o*-carboxyphenylamino)-1-deoxyribulose 5-phosphate (CdRP) into indole-3-glycerol phosphate, whereby carbon dioxide and water are released. The steady-state constants deduced from saturation curves for the substrate CdRP are listed in Table 3. All H2r mutant proteins show a much higher K_M^{CdRP} than both wild-type sIGPS and the control proteins, indicating that substrate binding is compromised. The k_{cat} value is unaltered for the control protein S58A and slightly increased for all other control and H2r mutant proteins besides S181A. The 7-fold decrease in the k_{cat} value caused by the S181A exchange indicates that the specific occupation of this position influences catalysis. As a consequence, catalytic efficiency k_{cat}/K_M^{CdRP} is notably lowered for four of the six H2r mutant proteins, namely, 13-fold (W8A), 15- or 20-fold (R54K/R64F and R54A/R64A), and 120-fold (S181A). In contrast, k_{cat}/K_M^{CdRP} is changed at most 2-fold for all control proteins (Table 3).

The stability of the sAnPRT and sIGPS H2r mutant and control proteins was analyzed by thermal unfolding of the secondary structure, which was monitored by the loss of the far-UV CD signal (Figures S1 and S2 of the Supporting Information). Because thermal denaturation was irreversible, the free energies of unfolding could not be determined. Instead, the apparent melting temperature at which 50% of the signal is lost (T_M^{app}) was used as an operational measure for stability. The sAnPRT and sIGPS wild-type proteins are extremely thermostable, showing T_M^{app} values of 91 and 92 °C, respectively (Tables 2 and 3). The stabilities of the sAnPRT H2r mutant proteins with the F297A and F297Y replacements and of the control proteins with the K203A, I316A, and K318A exchanges differ by at most 2 °C from that of wild-type sAnPRT. In contrast, the thermal stabilities of the other sAnPRT mutant proteins are compromised more drastically, albeit to a different extent. In the H2r mutant proteins, substitutions E50A, E54A, and I235A lead to a decrease of 4−6 °C, and the I235E and V284A exchanges reduce thermal stability by as much as 11 and 14 °C, respectively. In the control proteins, substitutions E123A and F251A lead to a decrease in the apparent melting temperature of 8−9 °C.

For sIGPS, the thermal stabilities of five H2r mutant proteins (W8A, S181A, S181N, R54A/R64A, and R54K/R64F) and three control proteins (S56A, S58A, and L187A) differ by not more than 2 °C from that of the wild-type enzyme. Two substitutions reduced T_M^{app} more significantly by 5 °C (H2r mutant protein W8K) and 4 °C (control protein V208A).

DISCUSSION

H2r Predicts Important Residue Positions with High Specificity but Limited Sensitivity. Methods analyzing correlated mutations in an MSA aim to identify residues that are crucial for the properties of a given protein family. The criterion used by H2r for such potentially important residue

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

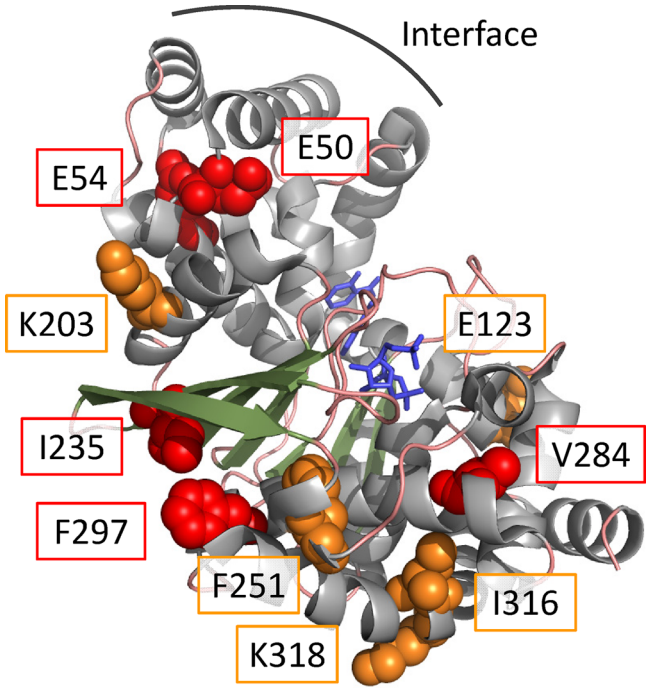


Figure 3. Putatively important residues of sAnPRT identified by H2r and control mutants. Ribbon presentation of one protomeric unit of homodimeric sAnPRT (PDB entry 1ZYK). Bound substrates (two molecules of anthranilate and one molecule of 5-phosphoribosyl-α-1-pyrophosphate, PRPP) are colored blue. The residues validated by mutagenesis are shown as spheres; those of the five H2r mutant proteins are colored red and those of the five control mutants orange. The interface region with the second protomer is indicated. α-Helices are colored gray, β-strands green, and loops pink. This figure was created with PyMOL.[46]

positions is the conn(k) value that corresponds to the number of other positions with which the considered position is strongly correlated (Figure 2). For an MSA of random composition and with default settings of H2r as used here, it is highly unlikely that a residue position gets assigned a conn(k) value of >3.[23] On the basis of an MSA of IGPS sequences, H2r identified 13 residue positions with a conn(k) value of >3. We assessed the two sIGPS residues W8 and S181, which have conn(k) values of 10 and 4, respectively, and are located close to the active site. In both cases, the k_{cat}/K_M^{CdRP} values of the respective alanine mutant proteins are severely compromised by a factor of 13 (W8A) or 120 (S181A). A 20-fold decrease in the catalytic efficiency was observed as a consequence of the R54A and R64A exchanges, which pertain to residues with conn(k) values of 7 and 14, respectively. In contrast, none of the control proteins, which contained amino acid exchanges at positions with proximity to the active site similar to that of the analyzed H2r mutant proteins, altered catalytic efficiency by a factor of more than 2. Whereas the H2r mutant protein W8K and the control protein V208A reduced thermal stability by −5 and −4 °C, respectively, all other exchanges do not compromise stability. The mean decrease in T_M^{app} values is 1.7 °C, both for the H2r mutant and for control proteins. Taken together, for positions located close to the active site of sIGPS, a noticeable effect on catalytic efficiency was observed only when replacing residues with a statistically significant conn(k) value.

For sAnPRT, H2r assigned a conn(k) value of >3 to not more than five positions, all of which were experimentally assessed. None of the H2r mutant proteins show k_{cat}/K_M^{PRPP} or k_{cat}/K_M^{AA} changes exceeding a factor of 3. In contrast, k_{cat}/K_M^{PRPP} and k_{cat}/K_M^{AA} values were moderately compromised by a factor of 2−8 in the control proteins. Thermal stability was decreased between 4 and 14 °C in five H2r mutant proteins and by 8 or 9 °C in two of the five control proteins. The mean loss of stability as a consequence of alanine replacements is 5.8 °C for H2r mutant proteins and 4.2 °C for control proteins. These findings indicate that for sAnPRT mean activity or stability effects observed when mutating positions with statistically significant conn(k) values generally do not exceed effects when control sites are mutated. However, the strongest observed destabilization by −14 °C is due to the V284A exchange, which pertains to the single position with the highest conn(k) value of 11. These findings are consistent with the anticipated inverse correlation between sensitivity and specificity notoriously afflicting all predictive algorithms. Concentrating on high conn(k) values gives rise to few predictions, which reliably indicate important residue positions. Reducing the cutoff for the conn(k) value increases the number of candidate sites but also the risk of false predictions, which means that the user has to find an optimal balance between coverage and effort devoted to the generation and characterization of a mutant protein. Along these lines, in an analysis restricted to sites with conn(k) values of >4, we would have missed the strongest activity effect, namely the 120-fold decrease in k_{cat}/K_M^{CdRP} caused by the S181A exchange in sIGPS, and the second strongest stability effect, namely the −11 °C decrease in T_M^{app} caused by the I235E substitution in sAnPRT.

Identification of Residues Leading to Beneficial Exchanges Is Still an Unsolved Problem. As with other methods used for correlation analysis, H2r provides information about neither the property (function or stability) for which a given position in the MSA is important nor the optimal residue to be introduced. In our study, this intrinsic limitation becomes most obvious from the observation that the S181A and S181N substitutions have inverse effects on the k_{cat} value of sIGPS (Table 3). This finding supports the conclusion that the identification of the most beneficial exchange based on an MSA is generally an unsolved problem. Along the same lines, our results confirm previous findings that the introduction of the consensus residue at a given position does not necessarily lead to an increase in protein stability.[40,42,43] In sIGPS, the consensus substitutions W8K and S181N decrease T_M^{app} by 5 and 2 °C, respectively (Table 3). Moreover, although the combined consensus exchanges R54K and R64K improve stability by 2 °C, this rather marginal improvement effect is offset by a 15-fold decrease in k_{cat}/K_M^{CdRP}. Furthermore, for sAnPRT, the consensus exchanges F297Y and I235E have no effect on stability or even drastically decrease T_M^{app} by −11 °C (Table 2).

The observation that conservation of a residue does not unambiguously signal a crucial role in catalysis or protein stability is further confirmed by the analysis of control proteins. The conservation of serine 56 or serine 58 in sIGPS is as high as 99.5 or 92.6%, respectively (Table 1). However, for both control proteins with the S56A and S58A exchanges, neither k_{cat}/K_M nor T_M^{app} is compromised. These findings confirm that each indicator and in silico method (including H2r) reaches a certain quality level when predicting crucial residues, which is not 100% because of the many evolutionary factors that are concurrently affecting the occupancy of residue positions.

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

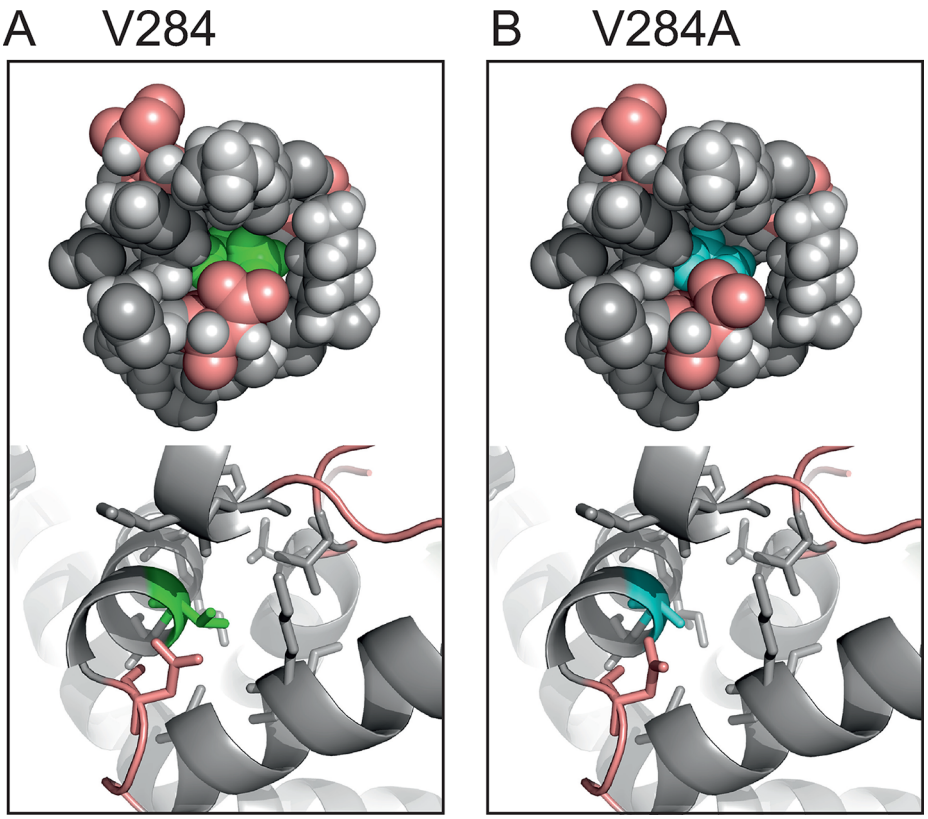


Figure 4. Local environment of V284 and V284A in sAnPRT. (A) Neighborhood in a 6 Å shell of V284 (green) as observed in the wild-type enzyme (PDB entry 1ZYK) in ball (top) and ribbon (bottom) representation; residue V284 is colored green, and α-helices are colored gray and loops pink. (B) Neighborhood in a 6 Å shell of the V284A exchange (cyan) modeled into the wild-type enzyme. The homology model was generated by means of FoldX 3.0;[47] PyMOL was used to create figures.[46]

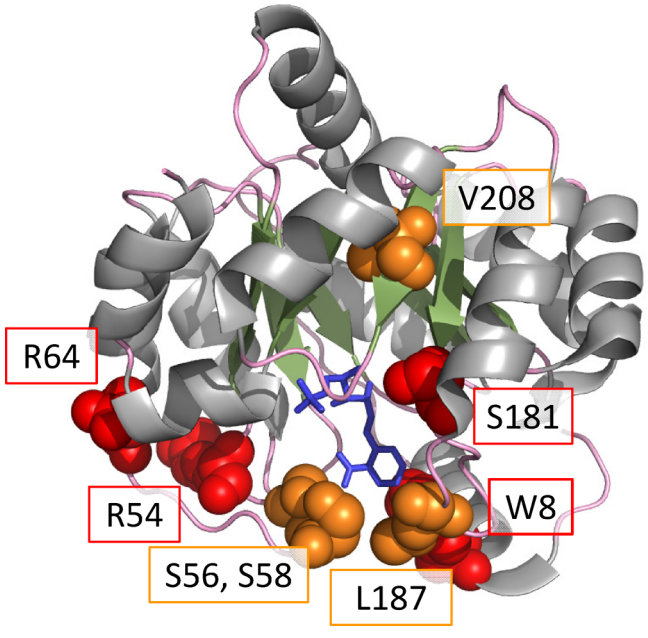


Figure 5. Putatively important residues of sIGPS identified by H2r and control mutants. Ribbon presentation of sIGPS (PDB entry 1LBF). The protein contains the bound substrate analogue [reduced 1-(*o*-carboxyphenylamino) 1-deoxyribulose 5-phosphate], which is colored blue. The residues tested by mutagenesis are shown as spheres; those of the four H2r mutant proteins are colored red and those of the four control mutants orange. α-Helices are colored gray, β-strands green, and loops pink. This figure was created with PyMOL.[46]

Structural Interpretation of Effects Caused by Replacing Residues Identified by H2r. Although H2r does not use any information other than amino acid sequence, available crystal structures for sAnPRT and sIGPS help to rationalize the consequences of the introduced amino acid exchanges for function and stability. The catalytic activities of the purified sAnPRT mutant proteins are largely unaltered, in accordance with the relatively large distance of the amino acid exchanges from the active site (Figure 3). In the absence of crystal structures for the mutant proteins, destabilization is difficult to explain. We therefore concentrated on the V284A replacement, which causes the strongest decrease in T_M^{app} of 14 °C (Table 2), in accordance with the largest conn(k) value of 11 (Table 1). Molecular modeling based on the crystal structure of wild-type sAnPRT suggests that the replacement of the valine side chain with the less bulky alanine side chain generates a small cavity within the interior of the protein (Figure 4). As a consequence, the protein core is less densely packed, explaining the strong destabilization caused by this exchange.[44,45] In agreement with this conclusion, the recently characterized T77I exchange of sAnPRT, which fills an internal cavity and results in denser packing, is strongly stabilizing.[40] The I235E exchange, which leads to the second most drastic decrease in stability [ΔT_M^{app} = −11 °C (Table 2)], might be caused by electrostatic repulsion with the neighboring E242 residue.

For sIGPS, the catalytic efficiencies of all H2r mutant proteins that introduce alanine are significantly decreased, mainly because of increased K_M^{CdRP} values (Table 3). This apparent reduction in substrate affinity can be explained by the

dx.doi.org/10.1021/bi300747r | *Biochemistry* 2012, 51, 5633−5641

spatial proximity of the exchanges, which are located close to the active site (Figure 5). Nevertheless, the results are not trivial as catalytic efficiency was not compromised by any of the alanine replacements in the control proteins, including the highly conserved serine residues 56 and 58. In this context, it is interesting to note that a comprehensive mutational analysis of the active site of IGPS from *E. coli* did not include residues corresponding to W8, R54, R64, and S181 of sIGPS, because of their lack of conservation.[41] The effects of the amino acid exchanges on the thermal stability of sIGPS are generally modest and difficult to explain in the absence of crystal structures for the mutant proteins.

Future Directions in Applying and Improving H2r. In a recent correlation analysis of enzyme superfamilies, mainly networks surrounding the active sites were identified, suggesting that the involved positions are more important for protein function than for structure and stability.[11] Our findings emphasize, however, that residue positions remote from the active site undergo correlated mutations as well and can be important for protein stability. These diverging findings are probably due to the different compositions of the studied MSAs. The MSAs compiled for the superfamilies comprised a large set of functionally diverse proteins, which can be divided into subfamilies. Most likely, the entire superfamily shares a set of correlated positions around the active site occupied by the substrate specificity-determining residues, which are conserved within the various subfamilies. Such conservation patterns cause strong correlation signals, which mask weaker signals elsewhere in the protein. In contrast, the MSAs compiled for an analysis by H2r do not contain highly similar sequences, which allowed us to identify less dominant correlations in AnPRT and IGPS. It is an attractive perspective to combine the two methods of correlation analysis to discriminate between positions being crucial for function and stability in large superfamilies of enzymes.

It is important for a computational analysis of MSAs to keep the number of identified candidate amino acids small enough to allow for their experimental validation with reasonable effort. An obvious selection criterion for mutational analysis is a high degree of conservation. Our results show that H2r extends the number of promising candidate residues beyond conserved ones as it predicts with a reasonable success rate crucial positions that show a high degree of sequence variability. When the MSAs are sorted according to degree of conservation, the positions identified by H2r rank between 33 and 309 for AnPRT and between 44 and 89 for IGPS (Table 1 and Tables S2 and S3 of the Supporting Information).

A prerequisite for an application of H2r is a sufficiently large MSA covering a broad range of sequence variability. After filtering, at least 125 sequences have to remain. However, this requirement is no longer a severe limitation for many proteins given that thousands of prokaryotic genomes have been sequenced over the past few years. This large data pool will allow for the further experimental validation of predictions made by H2r, and the outcome of these experiments shall lead to further improvements in the algorithm, e.g. with respect to sensitivity.

■ ASSOCIATED CONTENT

Supporting Information

Oligonucleotides used for site-directed mutagenesis of *strp*D and *strp*C (Table S1), conservation and conn(k) values for all residue positions of AnPRT proteins (Table S2), conservation and conn(k) values for all residue positions of IGPS proteins (Table S3), thermal denaturation of sAnPRT mutant proteins (Figure S1), and thermal denaturation of sIGPS mutant proteins (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-941-943 3015 (R.S.) and +49-941-943 3086 (R.M.). Fax: +49-941-943 2813. E-mail: reinhard.sterner@biologie.uni-regensburg.de (R.S.) and rainer.merkl@biologie.uni-regensburg.de (R.M.).

Author Contributions

S.D., N.B., and S.S. contributed equally to this work.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft within Priority Program SPP 1395 (STE 891/8-1 and ME 2259/1-1).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Sonja Fuchs, Jeanette Ückert, Kathleen Burghardt, Alexandra Perras, and Thomas Schwab for experimental assistance and discussion.

■ ABBREVIATIONS

AnPRT, anthranilate phosphoribosyl transferase; IGPS, indole-3-glycerol phosphate synthase; sAnPRT, AnPRT from *S. solfataricus*; sIGPS, IGPS from *S. solfataricus*; PDB, Protein Data Bank.

■ REFERENCES

(1) Capra, J. A., and Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics 23*, 1875−1882.

(2) Wang, K., and Samudrala, R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinf. 7*, 385.

(3) Amin, N., Liu, A. D., Ramer, S., Aehle, W., Meijer, D., Metin, M., Wong, S., Gualfetti, P., and Schellenberger, V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng., Des. Sel. 17*, 787−793.

(4) Lehmann, M., Loch, C., Middendorf, A., Studer, D., Lassen, S. F., Pasamontes, L., van Loon, A. P., and Wyss, M. (2002) The consensus concept for thermostability engineering of proteins: Further proof of concept. *Protein Eng. 15*, 403−411.

(5) Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins 18*, 309−317.

(6) Lockless, S. W., and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science 286*, 295−299.

(7) del Sol Mesa, A., Pazos, F., and Valencia, A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol. 326*, 1289−1302.

(8) Benkovic, S. J., and Hammes-Schiffer, S. (2003) A perspective on enzyme catalysis. *Science 301*, 1196−1202.

(9) Estabrook, R. A., Luo, J., Purdy, M. M., Sharma, V., Weakliem, P., Bruice, T. C., and Reich, N. O. (2005) Statistical coevolution analysis and molecular dynamics: Identification of amino acid pairs essential for catalysis. *Proc. Natl. Acad. Sci. U.S.A. 102*, 994−999.

(10) Marino Buslje, C., Teppa, E., Di Domenico, T., Delfino, J. M., and Nielsen, M. (2010) Networks of high mutual information define

the structural proximity of catalytic sites: Implications for catalytic residue identification. *PLoS Comput. Biol. 6*, e1000978.
(11) Kuipers, R. K., Joosten, H. J., Verwiel, E., Paans, S., Akerboom, J., van der Oost, J., Leferink, N. G., van Berkel, W. J., Vriend, G., and Schaap, P. J. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins 76*, 608–616.
(12) Kowarsch, A., Fuchs, A., Frishman, D., and Pagel, P. (2010) Correlated mutations: A hallmark of phenotypic amino acid substitutions. *PLoS Comput. Biol. 6*, e1000923.
(13) Magliery, T. J., Lavinder, J. J., and Sullivan, B. J. (2011) Protein stability by number: High-throughput and statistical approaches to one of protein science's most difficult problems. *Curr. Opin. Chem. Biol. 15*, 443–451.
(14) Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., and Ranganathan, R. (2005) Evolutionary information for specifying a protein fold. *Nature 437*, 512–518.
(15) Altschuh, D., Lesk, A. M., Bloomer, A. C., and Klug, A. (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol. 193*, 693–707.
(16) Neher, E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U.S.A. 91*, 98–102.
(17) Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. (2000) Correlations among amino acid sites in bHLH protein domains: An information theoretic analysis. *Mol. Biol. Evol. 17*, 164–178.
(18) Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics 21*, 4116–4124.
(19) Larson, S. M., Di Nardo, A. A., and Davidson, A. R. (2000) Analysis of covariation in an SH3 domain sequence alignment: Applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol. 303*, 433–446.
(20) Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol. 10*, 59–69.
(21) Dekker, J. P., Fodor, A., Aldrich, R. W., and Yellen, G. (2004) A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics 20*, 1565–1572.
(22) Tillier, E. R., and Lui, T. W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics 19*, 750–755.
(23) Merkl, R., and Zwick, M. (2007) H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinf. 9*, 151.
(24) Gao, H., Dou, Y., Yang, J., and Wang, J. (2011) New methods to measure residues coevolution in proteins. *BMC Bioinf. 12*, 206.
(25) Merz, A., Yee, M. C., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P., Yanofsky, C., and Kirschner, K. (2000) Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry 39*, 880–889.
(26) Schlee, S., Deuss, M., Bruning, M., Ivens, A., Schwab, T., Hellmann, N., Mayans, O., and Sterner, R. (2009) Activation of anthranilate phosphoribosyltransferase from *Sulfolobus solfataricus* by removal of magnesium inhibition and acceleration of product release. *Biochemistry 48*, 5199–5209.
(27) Mayans, O., Ivens, A., Nissen, L. J., Kirschner, K., and Wilmanns, M. (2002) Structural analysis of two enzymes catalysing reverse metabolic reactions implies common ancestry. *EMBO J. 21*, 3245–3254.
(28) Hennig, M., Darimont, B., Sterner, R., Kirschner, K., and Jansonius, J. N. (1995) 2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile *Sulfolobus solfataricus*: Possible determinants of protein stability. *Structure 3*, 1295–1306.
(29) Marino, M., Deuss, M., Svergun, D. I., Konarev, P. V., Sterner, R., and Mayans, O. (2006) Structural and mutational analysis of substrate complexation by anthranilate phosphoribosyltransferase from *Sulfolobus solfataricus*. *J. Biol. Chem. 281*, 21410–21421.
(30) Hennig, M., Darimont, B. D., Jansonius, J. N., and Kirschner, K. (2002) The catalytic mechanism of indole-3-glycerol phosphate synthase: Crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, substrate, and product. *J. Mol. Biol. 319*, 757–766.
(31) Schwab, T., Skegro, D., Mayans, O., and Sterner, R. (2008) A rationally designed monomeric variant of anthranilate phosphoribosyltransferase from *Sulfolobus solfataricus* is as active as the dimeric wild-type enzyme but less thermostable. *J. Mol. Biol. 376*, 506–516.
(32) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389–3402.
(33) Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009) NCBI Reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res. 37*, D32–D36.
(34) Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res. 33*, 511–518.
(35) Sarkar, G., and Sommer, S. S. (1990) The "megaprimer" method of site-directed mutagenesis. *BioTechniques 8*, 404–407.
(36) Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene 77*, 51–59.
(37) Jürgens, C., Strom, A., Wegener, D., Hettwer, S., Wilmanns, M., and Sterner, R. (2000) Directed evolution of a $(\beta\alpha)_8$-barrel enzyme to catalyze related reactions in two different metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A. 97*, 9925–9930.
(38) Schneider, B., Knöchel, T., Darimont, B., Hennig, M., Dietrich, S., Babinger, K., Kirschner, K., and Sterner, R. (2005) Role of the N-terminal extension of the $(\beta\alpha)_8$-barrel enzyme indole-3-glycerol phosphate synthase for its fold, stability, and catalytic activity. *Biochemistry 44*, 16405–16412.
(39) Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) How to measure and predict the molar absorption coefficient of a protein. *Protein Sci. 4*, 2411–2423.
(40) Schwab, T., and Sterner, R. (2011) Stabilization of a metabolic enzyme by library selection in *Thermus thermophilus*. *ChemBioChem 12*, 1581–1588.
(41) Darimont, B., Stehlin, C., Szadkowski, H., and Kirschner, K. (1998) Mutational analysis of the active site of indoleglycerol phosphate synthase from *Escherichia coli*. *Protein Sci. 7*, 1221–1232.
(42) Seitz, T., Thoma, R., Schoch, G. A., Stihle, M., Benz, J., D'Arcy, B., Wiget, A., Ruf, A., Hennig, M., and Sterner, R. (2010) Enhancing the stability and solubility of the glucocorticoid receptor ligand-binding domain by high-throughput library screening. *J. Mol. Biol. 403*, 562–577.
(43) Sullivan, B. J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., Syu, T., and Magliery, T. J. (2012) Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *J. Mol. Biol. 420*, 384–399.
(44) Ohmura, T., Ueda, T., Ootsuka, K., Saito, M., and Imoto, T. (2001) Stabilization of hen egg white lysozyme by a cavity-filling mutation. *Protein Sci. 10*, 313–320.
(45) Clark, A. T., McCrary, B. S., Edmondson, S. P., and Shriver, J. W. (2004) Thermodynamics of core hydrophobicity and packing in the hyperthermophile proteins Sac7d and Sso7d. *Biochemistry 43*, 2840–2853.
(46) *PyMOL* (2010) Schrödinger, Inc., Portland, OR.
(47) Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005) The FoldX web server: An online force field. *Nucleic Acids Res. 33*, W382–W388.