Published in final edited form as:

J Mol Biol. 2013 April 26; 425(8): 1378–1389. doi:10.1016/j.jmb.2013.01.034.

JANUS: Prediction and Ranking of Mutations Required for Functional Interconversion of Enzymes

Trevor A. Addington^{1,*}, Robert W. Mertz^{1,2,*}, Justin B. Siegel³, James M. Thompson³, Andrew J. Fisher^{1,4}, Vladimir Filkov⁵, Nicholas Fleischman¹, Alisa Suen¹, Chensong Zhang¹, and Michael D. Toney^{1,#}

¹Department of Chemistry, University of California, Davis CA 95616

²Biochemistry and Molecular Biology Graduate Group, University of California, Davis CA 95616

³Department of Biochemistry, Box 357350, 1705 NE Pacific St., University of Washington, Seattle WA 98195-7350

⁴Department of Molecular and Cellular Biology, University of California, Davis CA 95616

⁵Department of Computer Science, University of California, Davis CA 95616

Abstract

Identification of residues responsible for functional specificity in enzymes is a challenging and important problem in protein chemistry. Active site residues are generally easy to identify, but residues outside the active site are also important to catalysis and their identities and roles are more difficult to determine. We report a method based on analysis of multiple sequence alignments, embodied in our program Janus, for predicting mutations required to interconvert structurally related but functionally distinct enzymes. Conversion of aspartate aminotransferase into tyrosine aminotransferase is demonstrated and compared to previous efforts. Incorporation of 35 predicted mutations resulted in an enzyme with the desired substrate specificity but low catalytic activity. A single round of DNA back-shuffling with wild type aspartate aminotransferase on this variant generated mutants with tyrosine aminotransferase activities better than those previously realized from rational design or directed evolution. Methods such as this, coupled with computational modeling, may prove invaluable in furthering our understanding of enzyme catalysis and engineering.

Introduction

The ability to predict which residues in a protein determine a specific function (*e.g.*, substrate or reaction specificity, thermostability) is a long-standing and elusive problem in protein chemistry. One obstacle is that a compilation of diverse sequences, all of which have the same biological activity, includes a mixture of structural and functional information as well as evolutionary drift. For closely related enzymes of known structure, determinants of

© 2013 Elsevier Ltd. All rights reserved.

#Corresponding author: mdtoney@ucdavis.edu.

*These authors contributed equally.

Accession Numbers

PDB IDs for the structures reported here are: wild type AAT, 4F5M; J1, 4F5F; J2, 4F5G; J3, 4F5H; J4, 4F5I; J5, 4F5J; J6, 4F5K.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

specificity can sometimes be gleaned by active site comparisons, and confirmed by mutagenesis. However, a growing body of evidence suggests that residues distant from the active site also contribute to functional properties¹.

Probing distant residues for their effects on activity by site-directed mutagenesis (SDM) becomes impractical when many combinations of mutations require testing. If selection or screening for activity is possible, iterative rounds of directed evolution can be employed. However, only a small fraction of the sequence space can be sampled using current directed evolution methods, and many residues that may contribute to function may well be missed.

We report a bioinformatics method (named “Janus”) to identify and rank amino acid residues that specify differences in functionality between two structurally similar proteins. Janus combines analyses of structural conservation and mutation correlations, and is based on physicochemical properties of amino acids rather than amino acid identity alone. It is used here to predict the mutations required to convert one enzyme into a structurally related one of different substrate specificity, drastically reducing the exploration of sequence space required to interconvert enzyme function.

A variety of methods for identifying functional positions from sequence alignments of two or more groups of proteins have been reported (Evolutionary Trace², SDPpred³, GroupSim⁴, SCA⁵, *etc.*). Some methods, for example WebProAnalyst⁶, allow correlation of activity with physicochemical properties but do not compare two or more families or sequences. Analysis of correlated mutations, including correlated physicochemical properties, within protein families has also received significant attention (Chakrabarti and Panchenko⁷, CRASP⁸, SCA⁵). Most methods for comparing multiple families, with the notable exception of SCA, were primarily tested against benchmarks of existing data. To our knowledge Janus is the only method created with transferring functional properties between proteins in mind, and is validated experimentally here.

To calculate the functional importance of residues, Janus analyzes multiple sequence alignments (MSAs) of two structurally related proteins with distinct properties (referred to as “Start” and “Target” proteins), and includes structural information if available (see Supplementary Fig. 1 for an overview of the algorithm). Janus scores the alignments by measuring differences in conservation and covariation of *physicochemical properties* between the two sequence groups, and weights the scores according to their distance from the active site. The combined sequence and structural information is used to construct rank scores for residue-specific conservation (*ID*), and covariation (*CvD*). The *ID* and *CvD* scores, in addition to the predicted mutation required to impart Target functionality into the Start reference sequence, are output for all residues in the Start reference sequence that are not identical to the Target consensus amino acid.

Here, Janus is applied to the conversion of aspartate aminotransferase (AAT) into tyrosine aminotransferase (TAT) to evaluate the accuracy of its predictions. This is an excellent test case since the two enzymes are structurally homologous (Figure 1A; rmsd = 1.5 Å) but sequence divergent (43% identity), and AAT has very low levels of catalytic activity toward aromatic amino acids.

Despite the difference in catalytic properties, no absolutely conserved set of active site mutations distinguishes AAT from TAT, making the determinants of substrate specificity difficult to ascertain. Conversions of *E. coli* AAT into enzymes with TAT activity have been performed by both rational design and directed evolution (random mutagenesis)^{9; 10; 11; 12}. Using these methods, the best enzyme found had a k_{cat} value and substrate affinity significantly less than that of wild type TAT. Therefore, residues other than those previously identified are important for converting AAT into TAT.

Based on Janus predictions, we initially constructed a 35 mutation variant of AAT (J0) that has aromatic amino acid affinities close to those of TAT, but has a low k_{cat} (0.07% of TAT, Table 1). However, starting with J0, a single round of back-shuffling with wild type AAT resulted in enzymes with k_{cat} values greater than literature ones for TAT and k_{cat}/K_M values similar to those obtained previously. Janus successfully identified residues required for converting the substrate specificity of AAT into that of TAT. It should be generally applicable for identifying residues that engender functional differences between proteins.

Results

Computational prediction of functional residues

AAT and TAT sequences used as Janus input were compiled by Rothman and Kirsch¹⁰ and aligned with CLUSTAL¹³. An *E. coli* AAT structure (PDB ID 1ARG)¹⁴ was used for structural information. Janus output consists of two scores, conservation (*ID*) and covariation (*CvD*), per residue in the MSA and a mutation for each residue (original input, output, and program in Supplementary Materials). The *ID* score is a function of both the degree of conservation within each MSA, and the difference in conservation between MSAs at that position. The *CvD* score measures differences in correlated physical properties between MSAs, since large differences in correlations between groups are hypothesized here to indicate functional importance. Both scores are weighted by the inverse distance of the residue under consideration to the active site, since residues closer to the active site are generally expected to have a larger impact on function. A full description of the Janus algorithm, as well as the current version of the program, are provided as Supplementary Materials.

We examined the Janus rankings of residues identified in previous experiments and the overall distribution of scores to determine how many of the highest scoring predicted mutations to include in the initial construct. In previous experiments, twelve residues were identified as functionally important^{9; 10; 11; 12}: A12, P13, N34, V39, K41, T47, N69, T109, G261, S285, A293, and N297. Four of these residues (P13, V39, K41, and T47) were not considered in Janus since the TAT consensus amino acid is identical to *E. coli* AAT. Seven of the remaining eight are found in the 10 top *ID* and *CvD* scores, and the remaining residue (A293) is ranked 13th in *CvD*. The overall distribution of scores suggested to us that beyond the top 20–30 mutations for each score a baseline is reached and relative rankings are no longer functionally significant (Supplementary Figure 2). Therefore, the top 20 mutations from each score were incorporated into wild type AAT (Supplementary Table 1), resulting in 35 unique mutations since five occur in both conservation and covariation scores. This set includes three residues lining the substrate and cofactor binding pockets and 32 in the second and third shell surrounding the active site (Figure 2).

Synthetic gene and back-shuffling

A synthetic gene encoding *E. coli* AAT with the top 35 predicted mutations (J0) was procured from DNA2.0 (Menlo Park, CA) and subcloned into an expression vector for characterization. The protein was expressed and purified (Supplementary Methods), and its steady-state kinetic parameters were measured (Table 1). J0 has binding affinities for Phe and Tyr close to those for TAT (whereas wild type AAT shows no saturation with them), and a K_M for phenylpyruvate (PP) of 0.6 mM. However, the k_{cat} values for J0 are 1,000–2,000-fold lower than those for TAT with aromatic amino acids. Thus, the top 35 mutations gave the desired binding properties, but some of the mutations included were deleterious to catalysis. DNA shuffling¹⁵ between J0 and wild type AAT was performed to remove deleterious mutations. Since the total library size here is $\sim 10^{10}$, a genetic selection was required to sample a significant fraction of it.

The J0 back-shuffled libraries were transformed into the aminotransferase-deficient *E. coli* strain MG204¹⁶, and plated on selective medium to select for TAT activity. Cells harboring plasmid-borne TAT grew in two days, while those with plasmid-borne AAT grew in one week. Fifty clones that provided growth faster than wild type AAT were isolated and sequenced. These retained on average about half the mutations from J0, as expected based on the 1:1 ratio of J0 and wild type AAT DNA used in the back-shuffling.

The growth data and active clone sequences were analyzed statistically. Linear models have been successfully applied to assess the impact of individual mutations in directed evolution experiments¹⁷. Thus, we chose stepwise linear regression to estimate the effect of individual mutations on TAT activity. We used growth rates of the 50 active clones and wild type AAT as observations, with the variables being the presence or absence of the mutations (only 30 since several contiguous mutations were linked in all clones). We interpret a positive coefficient for a mutation as having a beneficial effect, a negative coefficient as deleterious, and mutations excluded from the model as having a small or inconsistent effect on activity. The linear model we generated suggests I33V/N34D, I73L, T109S, F221I, A229G, and V387I are beneficial; I9L, A12Q, G19S, A21N/D22E, I37V, L50M, A115G, S139T, S145A/V146I, I198A, F223I, Y256C, S285G, A290L/I291A, A293T, N297S, M333L, and M392L are neutral (excluded mutations); and N69T, I76L, V135I, G261S, I353V, and F362Y/S363L are deleterious (Supplementary Table 2). The neutral mutations S285G, A293T, and N297S, as well as deleterious mutations N69T and G261S, are notable since these residues were mutated in previous AAT variants. It is possible that their impact is context dependent or smaller than expected since these mutations were not tested individually in previous experiments. Only one beneficial mutation lines the substrate binding pocket (T109S), consistent with the idea that mutations beyond the active site have a significant impact on enzyme functionality.

The increase in growth rates achieved by back-shuffling with AAT confirmed that the 35 mutations included in J0 included deleterious ones. Although back-shuffling with the wild type gene proved efficient in this instance, it would be better to identify deleterious mutations *a priori* and remove them from the list. One way to achieve this might be to model computationally their effects on protein stability and/or ligand binding energy.

Using Rosetta¹⁸, we evaluated the predicted substrate binding energy and protein stability for the 50 active enzymes identified through back-shuffling, as well as 44 inactive enzymes that were also identified in back-shuffling. Protein sequences were modeled using the 1ARG structure as template with phospho-5'-pyridoxyltyrosine modeled in the active site. Full-atom energies for the entire protein and the ligand interface are depicted in box plots, split into active and inactive groups, in Figure 3. There is a significant difference in the distribution of whole-protein energies but very similar distributions for the ligand interface energy. This is confirmed by the two-sample Kolmogorov-Smirnov test, giving p-values of 4.54×10^{-5} and 0.632 for complete enzyme and ligand interface energies, respectively. This suggests that Janus could be productively followed up with modeling to eliminate at least some deleterious mutations, and that modeling provides orthogonal information to Janus. However, Rosetta in its current state could not discriminate between active and inactive sequences based solely on active site interactions.

Kinetic Characterization

Six of the fastest growing clones (referred to as J1–J6), each with 13–19 of the original mutations, were kinetically characterized (Table 1 and Supplementary Table 3). All beneficial mutations identified by linear regression are present in all six clones, with the exceptions of I73L (5 of 6 clones) and V387I (3 of 6 clones). Most deleterious mutations are found in two or fewer clones; however, N69T and I76L are surprisingly found in six and

five of these clones, respectively. The general trend is that back-shuffling dramatically improved k_{cat} relative to J0, with three variants having higher k_{cat} values for Tyr compared to TAT values (Supplementary Table 3). The improved k_{cat} values generally come at the expense of aromatic amino acid affinity, with K_M for Phe reaching as high as 85 mM.

Although K_M for aromatic amino acids increased relative to J0 for all back-shuffled mutants, K_M for PP (the aromatic substrate under selection) improved in all those characterized (K_M for PP as low as 0.15 mM). Importantly, the PP and Glu k_{cat}/K_M values, *the kinetic parameters under selective pressure during back-shuffling*, generally increased ~1000-fold over J0.

J1, the clone with the best overall kinetic parameters, has 13 mutations compared to wild type AAT. Compared to J0, back-shuffling improved the J1 k_{cat}/K_m values for amino acid substrates 60–280-fold, largely by restoring near wild-type k_{cat} values (Table 1). J1 has k_{cat}/K_m values for the dicarboxylic amino acids Asp and Glu similar to those of wild type TAT. Compared to the rationally designed HEX mutant and the evolved mutant 8-2B, J1 has ~10-fold weaker aromatic amino acid affinity, but exhibits up to 10-fold higher k_{cat} ^{10; 12; 19}. The k_{cat}/K_m values for J1 with aromatic amino acids are similar to those reported for HEX and 8-2B, and 20–70-fold lower than TAT. Most importantly, the J1 k_{cat}/K_M value for PP is 1400-fold greater than J0, and 60-fold greater than wild type AAT.

Structural characterization

The X-ray crystal structures of wild type AAT and J1–J6 were solved to provide a structural basis for understanding how the mutations influence substrate specificity (Supplementary Table 4). Each enzyme was crystallized in the unliganded, open conformation with PLP covalently bound to the active site lysine (K258). Overall, the structures of J1–J6 are very similar to that of wild type AAT, with a C α RMSD ranging from 0.575 to 0.723 Å (Supplementary Figure 3).

The largest structural differences are seen in the conformation of the small domain. In solution, the enzyme is in equilibrium between an open (substrate binding) and a closed (catalytically active) conformation, with the small domain rotating ~6° toward the large domain in the closed structure²⁰. The mutant conformations are all closer to the closed wild type conformation (average C α rmsd = 1.5 Å for J1–J6 versus wild type AAT closed conformation) than the open wild type structure solved here (C α rmsd = 1.7 Å for open conformation AAT versus AAT closed conformation, Supplementary Figure 4). A shift of the equilibrium to the closed conformation was also suggested to play a role in the expanded substrate specificity of the HEX mutant²¹.

The most significant differences at the cofactor and substrate binding pocket occur in the W140 loop (Figure 4), which participates in both PLP and substrate binding. In most mutants, the W140 loop is shifted further into the cofactor binding pocket, possibly due to the S145A mutation present in all of them since the Ser145 side-chain hydroxyl interacts with the main-chain carbonyl of Pro141 in wild type AAT. A rotation of the cofactor ring and the W140 side chain is also observed for some mutants. The conformation of the S139T side chain varies significantly relative to wild type, but this mutation is only present in two variants.

Although one expects changes in the size and/or shape of the active site to occur with altered substrate specificity, the kinetic parameters of the mutant enzymes do not clearly correlate with any single physical property of the active site. Of the three possible mutations lining the substrate binding pocket, two (I37V and T109S) are found in all six mutant structures

and one (N297S) is observed in two. From regression analysis, T109S is expected to have the most significant beneficial impact on TAT activity.

The most active mutant, J1, has only two mutations within 8 Å of the substrate binding site (I37V and T109S, Figure 5). Six more mutations are found in the second shell around the active site (8–16 Å): N34D, N69T, I73L, S145A, V146I, and V387I. The remaining mutations are beyond 16 Å from the active site: I33V, I76L, A115G, F221I, and A229G. All six regression-identified beneficial mutations are present in J1, as are two deleterious (N69T and I76L) and several neutral ones. The mutations in J1 do not drastically reshape the substrate binding pocket. I37V and T109S may function by expanding the volume of the pocket to accommodate aromatic substrates. T109S may also stabilize an alternative hydrogen bond network for aromatic substrate binding, similar to that observed in *Paracoccus denitrificans* TAT²². The remaining beneficial mutations are too far from the active site to assign functions. Hopefully, future studies will shed light on how these mutations enhance catalysis, possibly through active site dynamics and interacting networks of residues.

Discussion

The conversion of AAT into TAT was chosen as a proof-of-concept for the Janus algorithm, since previous results from rational design and random mutagenesis provide an important context. Janus dramatically reduced the effort required to find AAT mutants that complement a TAT deficient bacterium, and that are efficient catalysts *in vitro*. There are 194 amino acid differences between *E. coli* AAT and the consensus sequence for TAT. An exhaustive analysis of these mutations would require that 2^{194} (10^{58}) unique sequences be tested. By reducing the number of mutations under consideration from 194 to 35, the mutational sequence space is reduced to 2^{35} (10^{10}). This latter number is within the scope of current experimental techniques (*e.g.*, gene shuffling) for testing libraries.

The genetic selection employed here on back-shuffled mutants led to dramatically improved k_{cat} values, weaker affinity for aromatic amino acids, and slightly improved affinity for PP. This is logical since PP and Glu (the *in vivo* amino donor) are the substrates under selection. In this regard, back-shuffling and selection led to a 1400-fold increase in the efficiency of J1 over J0. The observed increases in k_{cat} are significant due to the apparent difficulty in achieving high turnover rates in enzyme design²³.

We sequenced a large number of clones and measured their relative fitness using growth rates on our selective medium. Regression analysis of these data allowed us to evaluate the relative strength of the conservation and covariation scores. Beneficial mutations identified by linear regression were largely predicted by the conservation score, while deleterious mutations were split evenly between conservation and covariation (Supplementary Table 2). This suggests that the conservation score is more powerful for identifying functional residues, but it is not clear how great the difference in predictive power is based on the present results.

It is also possible to evaluate the influence of the distance weight on both scoring methods by examining the unweighted conservation scores and covariation scores (I_i and Cv_i , respectively, see Janus output file in Supplementary Materials). The top 20 I_i scores would have kept 17 of the 20 mutations predicted by ID_i scores, including all the beneficial conservation score mutations from the regression analysis. However, the top 20 Cv_i scores would only retain four of the 20 CvD_i mutations, of which three were neutral and one deleterious in our regression analysis. Therefore, the conclusion from the present data is that the distance-weighted conservation is relatively insensitive to location in the protein, but the

covariation score is heavily influenced by distance from the active site. While these trends apply to the specific example presented here, more general conclusions require additional diverse applications of Janus to provide a broader basis.

The results presented here show that Janus can identify mutations, both within and beyond the active site, that contribute to enzyme functionality. The central hypothesis in Janus is: *conservation of chemical properties determines biological functionality, while dissimilarity in chemical properties at equivalent positions determines functional differences*. Janus extracts this information from sequence alignments, and weights it based on distance from the active site. No specifics regarding the catalytic mechanism are necessary, and computational modeling can provide orthogonal information that can be incorporated. Unlike current *de novo* enzyme design methods, Janus does not require construction of a theozyme to be grafted onto an existing enzyme scaffold²⁴. On the other hand, Janus is limited to functional properties that already exist in nature. Nevertheless, its utility is expected to encompass not only catalytic properties, but other properties of proteins such as thermostability that can be gleaned from MSAs of different families. Further development of sequence analysis tools such as Janus and their integration with molecular modeling should enable robust identification of functionally important residues. These methods will be important to understanding enzyme catalysis in general and, in particular, to deciphering how residues outside the active site participate in catalysis.

Methods

Mutation prediction and synthetic genes

The Janus algorithm is described in detail in Supplementary Materials. The version of the Perl script used to generate the predictions in this experiment (Janus16v6), along with the current version and input and output, are available as Supplementary Materials. AAT and TAT sequence sets were subsets of the sequences originally compiled by Rothman and Kirsch¹⁰. Residues 109 and 297 were required to be Thr and Asn, respectively, to be considered as an AAT. The AAT and TAT sequence sets were aligned with CLUSTAL X 1.83¹³ in Multiple Alignment Mode using the BLOSUM 30 substitution matrix, gap opening penalty = 10, gap extension penalty = 1. The RCSB PDB entry 1ARG¹⁴ was used as the reference structure with *E. coli* AAT as the reference sequence. Distances to the active site were measured using the CG atom of phospho-5'-pyridoxylaspartate as the active site center. Tables of linear correlation coefficients were generated for volume, polarity, and hydrophilicity on the CRASP server⁸ with default settings, gap threshold set to 10%, variability threshold 3.

The J0 protein sequence was designed by incorporating the top 20 predicted mutations from the Janus *ID* and *CvD* scores into wild type *E. coli* AAT, resulting in 35 unique mutations. A codon-optimized synthetic gene encoding J0 was procured from DNA2.0 (Menlo Park, CA). A wild type AAT gene with maximum nucleotide identity to J0 was also ordered to optimize DNA shuffling between J0 and wild type AAT.

Library construction and selection for TAT activity

The selection for TAT activity was based on the arabinose-inducible vector pBAD (Invitrogen) and the aminotransferase-deficient strain MG204¹⁶ (genotype *his23, proB, trpA-605, lacI3, lacZ118, gyrA, rpsL, Δ aspC::kanr, tyrB, ilvE, recA:tn10*). AAT, TAT, and J0 were subcloned into pBAD for selection and DNA shuffling experiments. Selective medium consisted of M9 salts with 15 g/L agar, 0.4% (w/v) glucose, 0.02% (w/v) L-arabinose, 100 mg/L ampicillin, 100 mg/L kanamycin, 15 mg/L tetracycline, and 50 mg/L of each of the 20 standard amino acids except Tyr and Phe. Untransformed MG204 incubated

at 37° C is unable to produce colonies on this medium. Colonies of MG204 transformed with pBAD-AAT appear in 7 days. MG204 transformed with pBAD-TAT grow colonies in 2 days. MG204 transformed with J0 is unable to produce colonies on selective medium.

A back-shuffled library of J0 and AAT was generated by *in vitro* DNA shuffling¹⁵ using a protocol adapted from Suenaga *et al*²⁵. Parent genes were PCR amplified from pBAD plasmid DNA using specific skew primers in order to reduce the background level of unshuffled parental sequences²⁶. Equimolar quantities of both PCR amplified parent genes were combined and digested with 0.6 U DNaseI (NEB) in 100 µL 100 mM Tris-HCl pH 7.4, 1 mM MnCl₂ for 30 minutes at 15° C, followed by DNaseI inactivation at 90° C for 10 minutes. Fragments of ~50 bp were purified from a 2.5% low melting agarose gel with a Takara Recochip (Takara Bio, Inc.) followed by removal of ethidium bromide by a Centri-Sep desalting column (Princeton Separations).

The purified fragments (10 µg total) were resuspended in a PCR mixture (200 µM each dNTP, 1× *Pfu* DNA polymerase reaction buffer) and 2.5 U *PfuTurbo* DNA polymerase (Stratagene) was added per 100 µL reaction mixture. The following thermocycler program was used for reassembly: 94° C for 5 min; 94° C for 1 min, 52° C for 1 min, and 72° C for 1 min + 5 s each cycle (45 cycles); and 72° C for 10 min. The reassembled product was amplified by conventional PCR using skew primers to select for products with at least one crossover. The amplified library and pBAD vector were digested with restriction enzymes *Xho*I and *Hind*III and ligated with T4 DNA ligase at a 10:1 insert to vector molar ratio.

Ligations were transformed into electrocompetent MG204 cells by electroporation using standard techniques²⁷. Cells were rescued after electroporation with SOC media without antibiotic for 1 hour at 37° C with shaking. Prior to plating the cells were pelleted at 1,000 *g* for 5 minutes, the supernatant aspirated, and the pellet gently resuspended in warm M9 minimal media with 0.4% (w/v) glucose to remove excess Phe and Tyr from the SOC media. For identification of active clones, transformants were plated on selective medium and incubated at 37° C for up to 10 days. Active clones were assigned growth times for the appearance of colonies at 24-hour intervals. For identification of inactive clones, transformants were initially plated on rich medium with antibiotics, then plasmid DNA was isolated and re-transformed in MG204 and plated on selective medium as described above. Clones from rich medium that were unable to grow on selective medium in 10 days were classified as inactive.

Protein expression, purification, and kinetic characterization

The D-hydroxyisocaproate dehydrogenase (DHDH) gene was cloned from *Lactobacillus casei* genomic DNA for use as a coupling enzyme for steady-state assays on phenylalanine and tyrosine due to its broad substrate specificity for hydrophobic α-ketoacids²⁸. *L. casei* was a gift from Jeff Broadbent (Utah State University). Genomic DNA was extracted using a GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich). The gene was amplified by PCR and cloned into pET-28 (Novagen) with a C-terminal His-tag for expression and purification from *E. coli* BL21(DE3). Cells were grown at 37° C until OD₆₀₀ = 0.08, when the temperature was shifted to 42° C to promote expression of chaperone proteins. At OD₆₀₀ = 0.6 the temperature was dropped to 22° C and expression induced with 0.7 mM IPTG. Cells were harvested after 20 hours, lysed by sonication, and DHDH was purified from the soluble fraction by Ni²⁺-affinity chromatography. The protein yield after purification was approximately 100 mg from 6 L of culture. Malate dehydrogenase (MDH) and glutamate dehydrogenase (GDH) were purchased from Calzyme Laboratories, Inc. for use as coupling enzymes for activity assays on aspartate and glutamate, respectively.

E. coli MG204(DE3) was prepared with a λ DE3 lysogenization kit (Novagen) for overexpression of J0. J0 was subcloned into pET-46 (Novagen) with an N-terminal His-tag and overexpressed in MG204(DE3). Cells were grown in LB with appropriate antibiotic at 37° C to an OD₆₀₀ of 0.6, induced with 1 mM IPTG for 6 hours, then harvested and lysed by sonication. J0 was purified from the soluble fraction on a Ni²⁺-affinity column by FPLC and eluted with a linear 10–250 mM imidazole gradient. Secondary purification was performed by FPLC on a FastQ anion exchange column (Pharmacia). Steady state assays were performed at 25° C in 100 mM HEPES pH 7.5, 100 mM KCl, 20 μ M pyridoxamine-5'-phosphate, 150 μ M NADH, and 2 μ M J0. 1 U/mL MDH was used for aspartate assays, 0.5 U/mL DHDH for tyrosine and phenylalanine assays, and 1 U/mL GDH for glutamate assays. Glutamate assays were supplemented with 10 mM ammonium acetate as cosubstrate for GDH. Reactions were followed by disappearance of NADH at 340 nm.

Wild type AAT and six of the fastest growing shuffled clones (J1–J6) were subcloned into pET-45 (Novagen) with an N-terminal His-tag and overexpressed in BL21(DE3). Cells were grown overnight in 1 L of M9 salt media supplemented with 2% (w/v) glucose, 10 g/L tryptone, 5 g/L yeast extract, 0.5 mg/L CaCl₂·2H₂O, 16.7 mg/L FeCl₃·6H₂O, 0.18 mg/L ZnSO₄·7H₂O, 0.16 mg/L CuSO₄·5H₂O, 0.15 mg/L MnSO₄·4H₂O, 0.18 mg/L CoCl₂·6H₂O, 20.1 mg/L EDTA, and 100 mg/L ampicillin. The next morning cells were pelleted and resuspended in pre-warmed LB medium and induced with 1 mM IPTG for 6 hours at 37° C. Cells were lysed by sonication and purified by Ni²⁺-affinity chromatography as described above. After affinity purification ammonium sulfate was added to 95% saturation as a concentration and secondary purification step. The ammonium sulfate precipitate was pelleted by centrifugation (4,000 g, 20 min., 4° C) and the supernatant removed by aspiration. The pellet was resuspended in a small volume of saturated ammonium sulfate for storage. Steady state assays were performed at 25° C in 100 mM TEA pH 7.5, 100 mM KCl, 20 μ M pyridoxal-5'-phosphate, 150 μ M NADH, and 1–10 nM enzyme. 1–10 U/mL MDH was used for aspartate assays, 0.5 U/mL DHDH for tyrosine and phenylalanine assays, and 1–10 U/mL GDH for glutamate assays. Glutamate assays were supplemented with 10 mM ammonium acetate as cosubstrate for GDH. Reactions were followed by disappearance of NADH at 340 nm.

Crystallization

His-tagged wild type AAT and J1–J6 were crystallized by hanging drop vapor diffusion. Proteins were concentrated to 10 mg/mL in 20 mM TEA pH 7.0, 100 mM KCl prior to crystallization. Drops of 5 μ L protein were mixed with 5 μ L precipitant solution and microseeded with crushed AAT crystals. AAT, J3, and J4 were crystallized with 0.1 M sodium acetate pH 4.6, 8% (w/v) PEG 4000; J6 was crystallized with 0.1 M sodium acetate pH 4.8, 8% (w/v) PEG 4000; and J1, J2, and J5 were crystallized with 0.1 M sodium acetate pH 5.0, 8% (w/v) PEG 4000. Drops were equilibrated for 2 days at 4° C. AAT, J3, J4, and J6 crystals were frozen with liquid nitrogen in 30% (v/v) 2-methyl-2,4-pentanediol, and J1, J2, and J5 crystals were frozen in 30% (v/v) ethylene glycol.

Data collection and refinement

Data for J1, J2, J4, J5, and J6 were collected in-house with a Rigaku RU-H3R rotating anode X-ray source and diffraction data collected on a Bruker Smart 6000 CCD. The remainder of the data were collected on beamline 9-2 at the Stanford Synchrotron Radiation Lightsource (SSRL). Diffraction data were processed with XDS. Crystal parameters, data collection, and refinement statistics are given in Supplementary Table 4. The structures were solved by molecular replacement using *E. coli* AAT (PDB ID 1ASE) as the search model. Model building was performed with COOT. Refinement, including TLS parameters, was accomplished with REFMAC followed by Phenix. Individual isotropic B-factors were

refined for structures refined at a resolution greater than 2 Å. Group occupancies were refined for residues with multiple conformations, as well as water, acetate, 2-methyl-2,4-pentandiol, and ethylene glycol molecules. TLS refinement used groups determined by Phenix. The PDB IDs for the structures reported here are: wild type AAT, 4F5M; J1, 4F5F; J2, 4F5G; J3, 4F5H; J4, 4F5I; J5, 4F5J; J6, 4F5K.

Regression analysis

Forward stepwise linear regression was performed with the MASS package in R^{29; 30}. Growth times for each active clone on selective medium were converted to growth rates by taking the inverse of the growth time in hours and normalizing to a TAT growth rate of one. Growth rates were used as the dependent variable, and each mutation term assigned a value of zero or one for wild-type or mutant, respectively. Five pairs of mutations contiguous in sequence were linked in all clones and are listed as single terms for analysis (*e.g.*, I33V/N34D), leaving 30 independent variables.

Rosetta modeling

Sequences of the active and inactive clones were modeled with Rosetta¹⁸ to examine structural differences between active and inactive enzymes. Alignments to template structures were generated using standard methods³¹. Full-atom models of protein structures were built using the standard Rosetta comparative modeling protocol³² with distance restraints designed to focus search near the template structures³³. We inferred details of ligand interactions by comparison to the phospho-5'-pyridoxyltyrosine complex modeled into the wild type AAT structure 1ARG. The interface between the ligand and modeled protein was jointly optimized using the Rosetta full-atom energy and catalytic restraints³⁴. For each protein 100 models were made, which were assembled into clusters by pairwise structural similarity, and the lowest energy model from the biggest cluster was defined as a representative structure for the sequence. For each representative sequence, the Rosetta full-atom ligand interface energy and full-atom energy was extracted and analyzed. The two-sample Kolmogorov-Smirnov test was performed in R.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grant GM54779 from the U.S. National Institutes of Health.

References

1. Lee J, Goodey NM. Catalytic Contributions from Remote Regions of Enzyme Structure. *Chemical Reviews*. 2011; 111:7595–7624. [PubMed: 21923192]
2. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*. 1996; 257:342–58. [PubMed: 8609628]
3. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucl Acids Res*. 2004; 32:W424–428. [PubMed: 15215423]
4. Capra JA, Singh M. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics (Oxford)*. 2008; 24:1473–1480.
5. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999; 286:295–299. [PubMed: 10514373]

6. Ivanisenko VA, Eroshkin AM, Kolchanov NA. WebProAnalyst: an interactive tool for analysis of quantitative structure-activity relationships in protein families. *Nucleic Acids Research*. 2005; 33:W99–W104. [PubMed: 15980590]
7. Chakrabarti S, Panchenko AR. Coevolution in defining the functional specificity. *Proteins-Structure Function and Bioinformatics*. 2009; 75:231–240.
8. Afonnikov DA, Kolchanov NA. CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences. *Nucl Acids Res*. 2004; 32:W64–68. [PubMed: 15215352]
9. Onuffer JJ, Kirsch JF. Redesign of the substrate specificity of *Escherichia coli* aspartate aminotransferase to that of *Escherichia coli* tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Sci*. 1995; 4:1750–7. [PubMed: 8528073]
10. Rothman SC, Kirsch JF. How does an enzyme evolved in vitro compare to naturally occurring homologs possessing the targeted function? Tyrosine aminotransferase from aspartate aminotransferase. *Journal of Molecular Biology*. 2003; 327:593–608. [PubMed: 12634055]
11. Chow MA, McElroy KE, Corbett KD, Berger JM, Kirsch JF. Narrowing substrate specificity in a directly evolved enzyme: the A293D mutant of aspartate aminotransferase. *Biochemistry*. 2004; 43:12780–7. [PubMed: 15461450]
12. Rothman SC, Voorhies M, Kirsch JF. Directed evolution relieves product inhibition and confers in vivo function to a rationally designed tyrosine aminotransferase. *Protein Sci*. 2004; 13:763–72. [PubMed: 14767072]
13. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL-W - IMPROVING THE SENSITIVITY OF PROGRESSIVE MULTIPLE SEQUENCE ALIGNMENT THROUGH SEQUENCE WEIGHTING, POSITION-SPECIFIC GAP PENALTIES AND WEIGHT MATRIX CHOICE. *Nucleic Acids Research*. 1994; 22:4673–4680. [PubMed: 7984417]
14. Graber R, Kasper P, Malashkevich VN, Sandmeier E, Berger P, Gehring H, Jansonius JN, Christen P. CHANGING THE REACTION SPECIFICITY OF A PYRIDOXAL-5'-PHOSPHATE-DEPENDENT ENZYME. *European Journal of Biochemistry*. 1995; 232:686–690. [PubMed: 7556224]
15. Stemmer WPC. RAPID EVOLUTION OF A PROTEIN IN-VITRO BY DNA SHUFFLING. *Nature*. 1994; 370:389–391. [PubMed: 8047147]
16. Fotheringham IG, Dacey SA, Taylor PP, Smith TJ, Hunter MG, Finlay ME, Primrose SB, Parker DM, Edwards RM. THE CLONING AND SEQUENCE-ANALYSIS OF THE ASPC AND TYRB GENES FROM *ESCHERICHIA-COLI-K12* - COMPARISON OF THE PRIMARY STRUCTURES OF THE ASPARTATE-AMINOTRANSFERASE AND AROMATIC AMINOTRANSFERASE OF *ESCHERICHIA-COLI* WITH THOSE OF THE PIG ASPARTATE-AMINOTRANSFERASE ISOENZYMES. *Biochemical Journal*. 1986; 234:593–604. [PubMed: 3521591]
17. Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C, Tam S, Muley S, Grate J, Gruber J, Whitman JC, Sheldon RA, Huisman GW. Improving catalytic function by ProSAR-driven enzyme evolution. *Nature Biotechnology*. 2007; 25:338–344.
18. Leaver-Fay, A.; Tyka, M.; Lewis, SM.; Lange, OF.; Thompson, J.; Jacak, R.; Kaufman, K.; Renfrew, PD.; Smith, CA.; Sheffler, W.; Davis, IW.; Cooper, S.; Treuille, A.; Mandell, DJ.; Richter, F.; Ban, YEA.; Fleishman, SJ.; Corn, JE.; Kim, DE.; Lyskov, S.; Berrondo, M.; Mentzer, S.; Popovic, Z.; Havranek, JJ.; Karanicolas, J.; Das, R.; Meiler, J.; Kortemme, T.; Gray, JJ.; Kuhlman, B.; Baker, D.; Bradley, P. ROSETTA3: AN OBJECT-ORIENTED SOFTWARE SUITE FOR THE SIMULATION AND DESIGN OF MACROMOLECULES. In: Johnson, ML.; Brand, L., editors. *Methods in Enzymology*. Vol. 487. Elsevier Academic Press Inc; San Diego: 2011. p. 545-574. Computer Methods, Pt C
19. Shaffer WA, Luong TN, Rothman SC, Kirsch JF. Quantitative chimeric analysis of six specificity determinants that differentiate *Escherichia coli* aspartate from tyrosine aminotransferase. *Protein Sci*. 2002; 11:2848–59. [PubMed: 12441383]
20. Jager J, Moser M, Sauder U, Jansonius JN. Crystal structures of *Escherichia coli* aspartate aminotransferase in two conformations. Comparison of an unliganded open and two liganded closed forms. *J Mol Biol*. 1994; 239:285–305. [PubMed: 8196059]

21. Malashkevich VN, Onuffer JJ, Kirsch JF, Jansonius JN. Alternating arginine-modulated substrate specificity in an engineered tyrosine aminotransferase. *Nat Struct Biol.* 1995; 2:548–53. [PubMed: 7664122]
22. Okamoto A, Nakai Y, Hayashi H, Hirotsu K, Kagamiyama H. Crystal structures of *Paracoccus denitrificans* aromatic amino acid aminotransferase: A substrate recognition site constructed by rearrangement of hydrogen bond network. *Journal of Molecular Biology.* 1998; 280:443–461. [PubMed: 9665848]
23. Baker D. An exciting but challenging road ahead for computational enzyme design. *Protein Science.* 2010; 19:1817–1819. [PubMed: 20717908]
24. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De Novo Enzyme Design Using Rosetta3. *PLoS One.* 2011; 6:e19230. [PubMed: 21603656]
25. Suenaga, H.; Goto, M.; Furukawa, K. *Evolutionary Methods in Biotechnology.* Wiley-VCH Verlag GmbH & Co. KGaA; 2004. DNA Shuffling; p. 13-24.
26. Ikeuchi A, Kawarasaki Y, Shinbata T, Yamane T. Chimeric gene library construction by a simple and highly versatile method using recombination-dependent exponential amplification. *Biotechnology Progress.* 2003; 19:1460–1467. [PubMed: 14524707]
27. Sambrook J, Russell DW. *Molecular cloning: A laboratory manual.* Molecular cloning: A laboratory manual. 2001
28. Hummel W, Schutte H, Kula MR. D-2-HYDROXYISOCAPROATE DEHYDROGENASE FROM *LACTOBACILLUS*-CASEI - A NEW ENZYME SUITABLE FOR STEREOSPECIFIC REDUCTION OF 2-KETOCARBOXYLIC ACIDS. *Applied Microbiology and Biotechnology.* 1985; 21:7–15.
29. Team, R. D. C. . *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; Vienna, Austria: 2012.
30. Venables, WN.; Ripley, BD. *Modern Applied Statistics with S.* 4. Springer; New York: 2002.
31. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics.* 2005; 21:951–960. [PubMed: 15531603]
32. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu GH, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D. NMR Structure Determination for Larger Proteins Using Backbone-Only Data. *Science.* 2010; 327:1014–1018. [PubMed: 20133520]
33. Thompson J, Baker D. Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins-Structure Function and Bioinformatics.* 2011; 79:2380–2388.
34. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. De Novo Enzyme Design Using Rosetta3. *Plos One.* 2011; 6:12.
35. Gloss LM, Kirsch JF. DECREASING THE BASICITY OF THE ACTIVE-SITE BASE, LYS-258, OF *ESCHERICHIA-COLI* ASPARTATE-AMINOTRANSFERASE BY REPLACEMENT WITH GAMMA-THIALYSINE. *Biochemistry.* 1995; 34:3990–3998. [PubMed: 7696264]
36. Luong TN, Kirsch JF. A general method for the quantitative analysis of functional chimeras: applications from site-directed mutagenesis and macromolecular association. *Protein Sci.* 2001; 10:581–91. [PubMed: 11344326]
37. Kuramitsu S, Hiromi K, Hayashi H, Morino Y, Kagamiyama H. PRE-STEADY-STATE KINETICS OF *ESCHERICHIA-COLI* ASPARTATE-AMINOTRANSFERASE CATALYZED-REACTIONS AND THERMODYNAMIC ASPECTS OF ITS SUBSTRATE-SPECIFICITY. *Biochemistry.* 1990; 29:5469–5476. [PubMed: 2201406]
38. Hayashi H, Inoue K, Nagata T, Kuramitsu S, Kagamiyama H. *Escherichia coli* aromatic amino acid aminotransferase: characterization and comparison with aspartate aminotransferase. *Biochemistry.* 1993; 32:12229–39. [PubMed: 8218300]

Highlights

- Introduces new computational method for identifying functionally relevant residues
- Janus greatly reduces sequence space needed for conversion of enzyme function
- Structures of highly active mutants provide examples of redesigned enzymes
- Computational modeling shown to have potential for screening deleterious mutations

A **B** **Figure 1.**

A. Structural similarity between AAT and TAT. *E. coli* AAT (grey, this work) and TAT (blue, PDB ID 3TAT) backbone traces are superposed with cofactor shown in space filling. The two proteins have 43% sequence identity and a C α RMSD of 1.15 Å. Native α_2 dimers are shown for each. **B.** Reactions catalyzed by AAT and TAT. AAT transfers the amino group between dicarboxylic amino acids and their corresponding α -keto acids. TAT catalyzes the same reaction but with expanded substrate specificity that includes aromatic α -amino and α -keto acids.

Figure 2. Distribution of Janus mutations on AAT. Wild type AAT (PDB ID 1ARG) is shown with the cofactor-substrate analogue in space filling. Residues mutated in J0 are colored blue with side-chains depicted as sticks to illustrate the distribution of predicted mutations with respect to the active site.

Figure 3.

Rosetta energies for active and inactive enzymes. Box plots of Rosetta full-atom energies for the full protein (left) and ligand interface (right) are displayed, each split between active and inactive enzymes. Whiskers represent maximal and minimal values. The distribution differs significantly between active and inactive enzymes for the full protein but not for the ligand interface. The two sample Kolmogorov-Smirnov test p-values are 4.54×10^{-5} and 0.632 for the full protein and ligand interface, respectively.

Figure 4. Comparison of wild type and mutant active site structures. Mutant structures J1–J6 (grey) superposed on wild type AAT (green). The loop that contains W140, which π stacks with PLP and is involved in hydrogen bonding to the α -carboxylate of the substrate, shows the greatest variation.

Figure 5. Mutations surrounding the J1 active site. The active site from the J1 Rosetta model with bound 5'-phosphopyridoxyltyrosine is shown. The 13 mutations are highlighted in blue with side chains displayed as sticks. Only I37V and T109S line the substrate binding pocket in the closed conformation.

Table 1

Kinetic constants for wild type and engineered enzymes

Enzyme	Asp			Phe			Tyr			PP		
	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($M^{-1} s^{-1}$) $\times 10^{-3}$	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($M^{-1} s^{-1}$) $\times 10^{-3}$	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($M^{-1} s^{-1}$) $\times 10^{-3}$	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($M^{-1} s^{-1}$) $\times 10^{-3}$
AAT	160 ^b	1.8 ^b	91 ^b	--	--	0.12 ^c	--	--	0.55 ^d	--	--	8.4 ^{e,f}
J0	0.27 (0.01)	1.8 (0.1)	0.15 (0.01)	0.24 (0.02)	1.2 (0.1)	0.20 (0.02)	0.14 (0.01)	0.7 (0.1)	0.21 (0.03)	0.21 (0.01)	0.62 (0.05)	0.34 (0.03)
J1	51 (1)	1.2 (0.1)	43 (4)	150 (10)	12 (1)	13 (1)	240 (10)	8 (1)	29 (3)	98 (5)	0.21 (0.03)	470 (70)
TAT	140 ^g	3.8 ^g	37 ^g	250 ^g	0.26 ^g	960 ^g	210 ^g	0.32 ^g	660 ^g	--	--	20 000 ^{g,f}
HEX ^h	7.4	0.22	34	29	0.78	37	24	0.65	37			
8-2B ⁱ	41	0.38	110	45	1.4	32	63	0.88	72			

Each amino acid substrate was reacted with α KG as amino acceptor. PP was paired with L-Glu as amino donor, which was the reaction under selection. Reaction conditions for J1: 100 mM TEA pH 7.5, 100 mM KCl, 20 μ M PLP, 150 μ M NADH. Reaction conditions for J0: 100 mM HEPES pH 7.5, 100 mM KCl, 20 μ M PMP, 150 μ M NADH. L-Asp was assayed with 1–10 U/mL MDH, L-Phe and L-Tyr with 0.5 U/mL DHDH, and PP with 10 U/mL GDH. 10 mM ammonium acetate was included in PP assays as cosubstrate for GDH.

^aGrowth times are time required for appearance of visible colonies on TAT selective media in hours or NG (no growth).

^bGloss & Kirsch³⁵.

^cLuong & Kirsch³⁶.

^dRothman & Kirsch¹⁰.

^eKuramitsu *et al*²⁷.

^fFrom half-reaction kinetics with PMP form of enzyme.

^gHayashi *et al*³⁸

^hRationally designed AAT hexamutant⁹; 12; 19.

ⁱDirectly evolved AAT mutant¹⁰.