Direct Coupling Analysis improves the identification of beneficial amino acid mutations for the functional thermostabilization of a delicate decarboxylase

Manfred Maier[1], Jan Esch[2], Alexander Schug[3] and Kersten S. Rabe[1,*]

[1]Institute for Biological Interfaces I, Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen, Germany

[2]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen, Germany

[3]Institute for Advanced Simulation, Jülich Supercomputing Center, 52428 Jülich, Germany

[*] correspondence should go to: Kersten Rabe, Institute for Biological Interfaces I, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany. Fax: + 49 (0)721 608 2-5546, Tel: + 49 (0)721 608 2-4496. kersten.rabe@kit.edu

Running title:
Thermostabilization via Direct Coupling Analysis

Abstract

The optimization of enzyme properties for specific reaction conditions enables their tailored use in biotechnology. Predictions using established computer-based methods, however, remain challenging, especially regarding physical parameters such as thermostability without concurrent loss of activity. Employing established computational methods such as energy calculations using FoldX can lead to the identification of beneficial single amino acid substitutions for the thermostabilization of enzymes. However, these methods require a 3D-structure of the enzyme. In contrast, coevolutionary analysis is a computational method, which is solely based on sequence data. To enable a comparison we employed coevolutionary analysis together with structure-based approaches to identify mutations, which stabilize an enzyme while retaining its activity. As an example, we used the delicate dimeric, thiamine pyrophosphate dependent enzyme ketoisovalerate decarboxylase and experimentally determined enzymatic activity and stability. Coevolutionary analysis led to the identification of beneficial mutations, which were not identified by other methods used in this study and had an overall success rate above 30%.

Keywords

Coevolutionary analysis, computational design, enzyme engineering, functional enzyme, thermostability,

Introduction

Proteins are excellent biocatalysts that are frequently used in industrial biotechnology (Zamost *et al.*, 1991). In order to develop efficient and new industrial bioprocesses, it is often desirable to improve the properties of naturally occurring enzymes. Especially the sensitivity to elevated temperatures often limits their industrial application potential (Bornscheuer *et al.*, 2012; Kristjansson, 1989). For example we recently established the on-demand production of reactor cartridges for flow chemistry applications by temperature-controlled 3D printing of enzyme containing gel-matrices which requires thermostable enzymes (Maier *et al.*, 2018). Apart from *in vitro* applications using purified enzymes, whole-cell based processes using mesophile and especially thermophile host organisms also require stable and robust enzymes for heterologous expression. In particular, the use of thermophile host organisms would enable high temperature fermentations that could have significant improvements compared to existing fermentations at moderate temperatures. The thermodynamics and kinetics of the reactions catalyzed are positively affected using increased temperatures, as well as reduced viscosity of complex nutrient media and a possible decrease in cooling costs when employing higher process temperatures (Abdel-Banat *et al.*, 2010). For example the production of isobutanol using the thermophile host organism *Geobacillus thermoglucosidasius* was described (Lin *et al.*, 2014). However, the restricted thermostability of the heterologously expressed mesophile key enzyme in the isobutanol pathway (α-ketoisovalerate decarboxylase, Kivd) was proposed to be limiting the yield of the fermentation. To solve such problems either naturally occurring thermostable enzymes can be identified or protein engineering can be used to stabilize an enzyme while retaining its activity (Buss *et al.*, 2018).

Since no thermostable homologue of KivD with comparable catalytic turnover has been described so far, we and others(Maier *et al.*, 2018; Soh *et al.*, 2017; Sutiono *et al.*, 2018) have employed rational and random mutagenesis resulting in significant thermostabilizations of up to 6,3°C due to a single amino acid mutation. Taking into account that most single amino acid substitutions only lead to small improvements along the evolutionary fitness landscape of an enzyme (Romero and Arnold, 2009), these stabilizations are significant. Testing huge libraries of protein mutants for functional thermostabilization

is only feasible using either a screen which couples the enzymes activity to the cells' survival (Schwab and Sterner, 2011) or which usually offers optical readouts (Xiao *et al.*, 2015). If no such screen is available, the number of all possible single amino acid substitutions in a protein usually exceeds experimental testing capacities since the production and analysis of individual variants of a protein is time-consuming and resource-intensive. Therefore numerous computer-aided methods have been developed that can guide the laboratory work with model-based predictions.

Common methods for stabilizing proteins *in silico* are either based on specific structural modifications like introducing disulfide bridges, helix dipole stabilization and hydrophobic core-repacking (Turner *et al.*, 2007) or calculations to minimize an effective energy mimicking the Gibbs free energy (Modarres *et al.*, 2016). The energetic relationships of atoms and molecules are approximated to maintain computational feasibility. A particular challenge is estimating conformational entropy changes and physical movements in the molecule in order to draw conclusions concerning its thermal stability (Zeiske *et al.*, 2016). In general, these calculations require an increase of computational resources with increasing protein size and precision. Therefore, static energy-based methods such as FoldX (Schymkowitz *et al.*, 2005) and Rosetta (Song *et al.*, 2013) have been developed, which do not simulate physical movements but instead calculate an effective energy-score based on a particular conformation also comparison entropic components. This is more efficient but the necessary simplification of complex interactions of the atoms within the protein and with the surrounding solvent limit the accuracy of these methods (Potapov *et al.*, 2009). However, new approaches combining standard and temperature-dependent statistical potentials with models trained on large data sets might further increase the predictions (Pucci *et al.*, 2016).

Although enzymes are three-dimensional machineries and their physical and catalytic properties are based on complex interactions within the structure, the information about these characteristics -in principle- is encoded in the amino acid sequence. Since the available genomic sequence information is growing much faster than structural or biochemical data correlated with the sequences, using primary amino acid sequence data to predict biochemical or biophysical parameters is a very

attractive field of research. Thus, apart from the above described structure-based approaches, statistical methods have been developed, that can make predictions based solely on sequence data such as provided by large multiple sequence alignments for a protein family. One such method is Direct Coupling Analysis (DCA) (Morcos *et al.*, 2011; Weigt *et al.*, 2009).

In contrast to the established analysis of the conservation of individual residues or amino acid pairs without considering other sequence positions ("local"), DCA uses a statistical model that considers the coevolution of all sequence positions pairs simultaneously. While computationally considerable more expensive than, e.g. simple mutual information, DCA has found widespread application in protein structure prediction providing accurate prediction of amino acid pair adjacencies (Dago *et al.*, 2012; Schug *et al.*, 2009; Uguzzoni *et al.*, 2017). Effectively, DCA estimates a fitness landscapes for sequences of a particular protein family. One striking example by M. Weigt and coworkers was using DCA to estimate the effect of single point mutations on the antibiotic resistance of ß-lactamase-producing bacteria (Figliuzzi *et al.*, 2016). They could demonstrate a positive correlation between experimentally determined antibiotic resistance and DCA energy of the respective lactamase. In a different study, coevolving mutation hotspots were identified to optimize the thermostabilization of an α-Amylase via saturation mutagenesis at these sites (Wang *et al.*, 2012).
The aim of this study was to investigate, whether DCA-analysis can improve rational protein approaches for the thermostabilization of a target enzyme. Thus, DCA was employed to predict and experimentally verify stabilizing mutations for an α-ketoisovalerate decarboxylase. This data was then compared to the results of previous evolution experiments (Maier *et al.*, 2018) classifying all mutations using DCA-analysis, the established FoldX algorithm and the HoTMuSiC webserver.

Results

As a test case for comparing different rational approaches for functional thermostabilization, we chose the enzyme a-ketoisovalerate decarboxylase from *Lactococcus lactis* (Uniprot accession number: A0A0BBQZ66, Figure 1A). The enzyme is a homodimeric protein with two active sites at the monomer-monomer interface containing each a non-covalent thiaminpyrophosphate

cofactor with no comparably active thermostable homologue described.

In addition to previously described methods, we employed DCA for the prediction of functional thermostabilized enzyme variants. DCA-analysis was originally developed to analyze amino acid contacts within the 3D structure of an enzyme in order to support structure predictions (Morcos *et al.*, 2011; Schug *et al.*, 2009; Weigt *et al.*, 2009). Furthermore, it can also be taken as a measure for the evolutionary fitness of a particular amino acid sequence. One component to this fitness should be the (thermo)-stability of proteins, since an increase stability should make the protein generally more robust to environmental conditions. To this end, we identified homologous sequences within the uniprot database and aligned them pairwise using the PHMMER tool (Finn *et al.*, 2015). In total 1679 homologous sequences of TPP-dependent enzymes were found, all showing three similar protein domains (N-terminal TPP-binding domain, central domain, C-terminal TPP-binding domain, Figure 1B), that were recognized using the Pfam-database (El-Gebali *et al.*, 2018) .

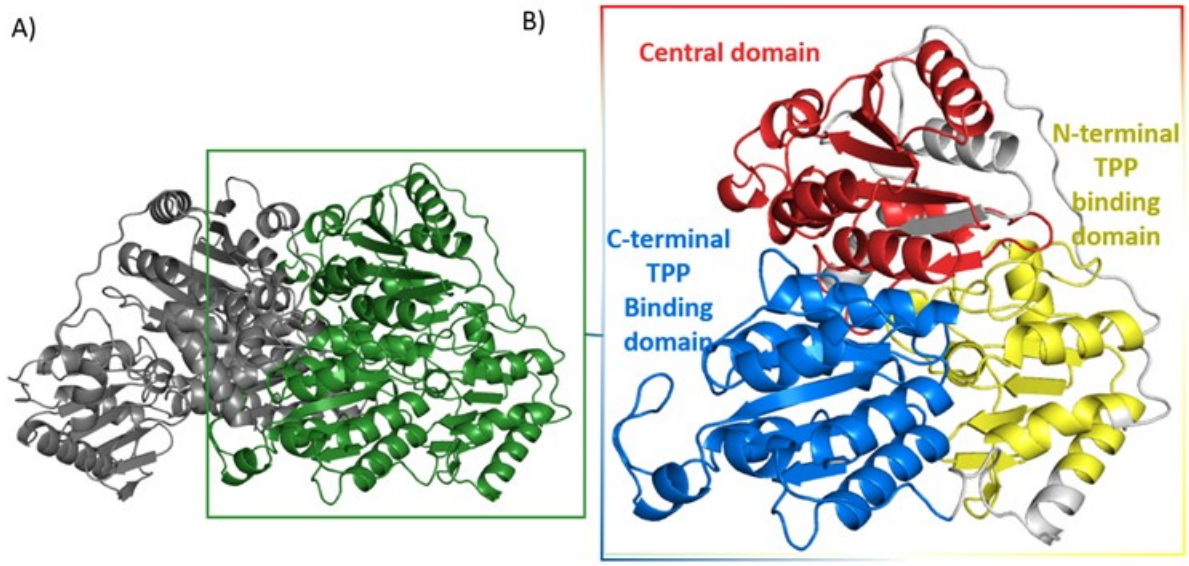

Figure 1: A) Model of the dimeric target enzyme Kivd, based on the KdcA structure (Berthold *et al.*, 2007). B) Domain architecture of a Kivd monomer recognized using the Pfam-database (El-Gebali *et al.*, 2018).

To reduce the effects of phylogenetic biases in the databases for DCA analysis, all sequences with a sequence identity higher than 80% were downweighed to reduce effects from phylogenetically closely related sequences (Schug *et al.*, 2009; Weigt *et al.*, 2009).

In Figure 2, the DCA results for all 10960 possible single amino acid substitutions (including a gap) are shown, with negative values indicating beneficial mutations. For the vast majority of mutations (98,5%) a positive DCA energy was calculated suggesting a deleterious impact on the evolutionary fitness of the amino acid sequence. Nevertheless, 168 mutations showed a negative DCA energy, indicating to be beneficial (Figure 2A, green dots). These mutations were distributed between 56 positions of the amino acid sequence, thus showing that the potential positions for beneficial mutations according to the DCA calculation seem to be clustered (Figure 2B).

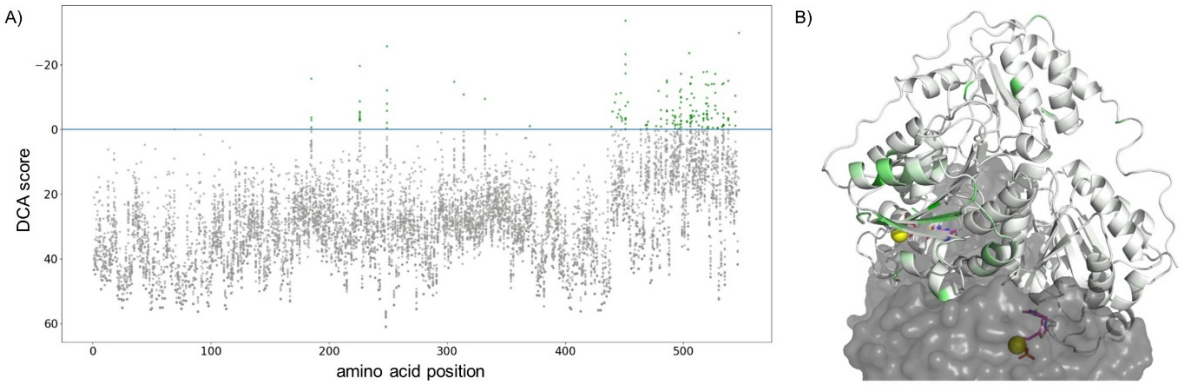

Figure 2: Graphical visualization of the DCA energy calculated for each possible mutation including a gap at each position of the amino acid sequence. A) Green dots represent negative DCA energies. Note that only 168 single mutation (1,5%) are predicted to have beneficial effects on the protein stability. B) Positions with advantageous mutations according to the DCA marked in green in one monomer of the enzyme. The color strength is scaled according to the DCA value. The second half of the dimer is shown as a grey surface.

Most of the potentially beneficial mutations are located within the C-terminal TPP-binding domain. In the cofactor binding pocket and the active site, no potentially beneficial mutations could be identified. This corresponds to the fact that the active site is the most evolutionarily conserved part of a protein (Das and Gerstein, 2004). Among the 25 most promising mutations according to their DCA energy, single amino acid substitutions for the experimental characterization were selected (Table 1). Since we were aiming to identify mutations which cannot be found by other established methods, we rejected all suggestions that correspond to the consensus sequence of the alignment. Furthermore, some mutations had already been investigated by us as a result of other directed evolution approaches. Also substitutions into amino acids with similar size and hydrophobicity (Kyte and Doolittle, 1982) were not selected for further analysis. Using these criteria eight amino acid

substitutions were selected, overexpressed and purified and their the T_{50}-value was analyzed. This was defined as the temperature at which 50% of the enzymes' initial activity is lost due to thermal denaturation after 10 min of incubation. From these selected mutations three led to less stable enzymes or no activity could be determined in our setup. However, three of the remaining mutations significantly stabilized the protein, increasing its T_{50} for up to 3,9°C. Although the number of protein variants characterized is small, this corresponds to a success rate of 37,5%, which is promising compared to established force-field based analysis (approx. 20% success rate)(Buss *et al.*, 2018) as well as random mutagenesis (2% success rate)(Broom *et al.*, 2017).

Table 1: Mutations selected from the top hits from the DCA calculation and their corresponding experimentally determined ΔT_{50}. Analysis was carried out at least in duplicate.

Entry	DCA energy	Consensus	selection criterium	Experimental ΔT_{50} [°C]
C451I	-33,58	I	T50 known: +2,1°C + consensus	
K547-	-29,81	-	End of sequence + consensus	
A249T	-25,68	T	best DCA energy	inactive
F505L	-23,56	L	consensus	
C451V	-23,2	I	T50 known: +1,5°C	
C451-	-20,05	I	three other mutations already studied	
S226T	-19,57	H		- 3,5 ± 0,4
Y520Q	-17,82	Q	consensus	
N517D	-17,64	D	consensus	
C451L	-17,25	I	T50 known: -6,5°C	
I498K	-17,14	K	consensus	
A527D	-17,09	P		+ 0,5 ± 0,3
S507A	-16,1	E		+ 1,2 ± 1,6
S185P	-15,67	P	consensus	
S486A	-15,06	A	consensus	
M538L	-14,98	-		+ 3,7 ± 0,3
D306N	-14,66	D	similiar amino acid	
S486V	-14,51	A	higher ranking mutation S486A	
V534L	-14,22	A		+ 3,9 ± 0,8
V506E	-14,22	E	consensus	
L526M	-14,03	M	consensus	
V506A	-14	E	higher ranking mutation V506E	
L524V	-13,99	V	consensus	
K533E	-13,91	K	similiar amino acid	
M519L	-13,86	L	consensus	
W521L	-13,66	F		inactive
A249D	-7,9	T		+ 1,6 ± 0,5

In previously published work, we utilized the FoldX algorithm to predict thermostabilizing mutations and found that approx. 20% of the resulting protein variants showed an increased stability while retaining the catalytic activity (Maier *et al.*, 2018). By combining the data, we purified and analyzed 31 individual Kivd variants. As a further bioinformatics method for the prediction of mutations, we used the HoTMuSiC webserver (Pucci *et al.*, 2016).

In Table 2 the ΔT_{50}-value compared to the wildtype of all characterized enzyme variants based on different selection strategies is listed. Additionally the predicted effect of these mutations according to different computational methods are shown, in order to compare the accuracy of these methods. Negative ΔΔG energies represent a lowering of the Gibbs' free energy of the enzyme and are thus considered to have a stabilizing effect. Using the HoTMuSiC webserver, the change in the melting temperature (TM) according to the respective mutation was predicted (Pucci *et al.*, 2016).

Table 2: Experimentally determined T_{50} and the corresponding DCA energy, ΔΔG calculated using FoldX and the ΔTm generated by the HoTMuSiC Server. Negative DCA energies as well as negative ΔΔG indicate potentially stabilizing mutations. Cases where the prediction correlates with the experimental data are marked. Green indicates a successful prediction of a stabilizing mutation, gray indicates a correct destabilizing prediction.*=Mutations selected based on the DCA calculations.

Mutation	Experimental ΔT_{50} [°C]	DCA energy	calculated ΔΔG [kcal/mol]	Calculated ΔT_m [°C] *HoTMuSiC*
S385M	+6.3	+23.44	-6.69	-1.20
V534L*	+3.9	-14.20	-2.60	-0.37
M538L*	+3.7	-15.00	+1.20	-1.02
C451I	+2.1	-33.85	-6.84	-0.71
A249D*	+1.6	-7.90	+2.0	+0.19
C451V	+1.5	-23.20	-5.43	-0.77
S507A*	+1.2	-16.10	-1.70	+0.07
S314V	1	+3.65	-0.69	-0.10
A527D*	+0.5	-17.10	+2.10	+0.31
S244G	+0.4	+9.84	-3.62	-0.03
K309I	+0.4	+25.53	-1.01	+0.40
K231E	-0.4	+24.27	+1.07	-0.43
A249S	-1.5	-12.00	+2.05	+0.60
G45P	-1.8	+44.68	-5.31	+0.87
S314R	-1.9	-10.76	-1.62	-0.28
E332G	-2.7	-9.37	-2.71	+0.75

S245A	-2.8	+16.80	-1.90	-2.0
S226T*	-3.5	-19.60	-1.10	-0.41
C451A	-3.6	+0.00	+1.64	-2.35
S185P	-4.3	-15.67	+1.47	-0.71
N240P	-6.0	+13.17	-5.83	-1.04
C451M	-6.2	-7.22	-4.59	-1.77
C451L	-6.5	-17.25	-1.73	-1.60
T101P	Inactive	+28.66	-6.04	-0.10
A152Q	Inactive	+19.08	-0.26	-0.77
D191L	Inactive	+24.31	-6.95	+2.19
A249T*	Inactive	-25.70	+2.04	+0.37
T367L	Inactive	+2.70	-1.45	-0.41
L441R	Inactive	+28.01	+8.13	-2.70
N479R	inactive	+0.48	-2.66	-0.61
W521L*	inactive	-13.70	+8.50	-3.06

Overall, the three methods perform comparably when predicting the trend of mutations (stabilizing and destabilizing, marked green or gray in Table 2, respectively). DCA predicted 17 out of the 31 experimental cases in this study correctly, FoldX 15 and HoTMuSiC 19. However, when only comparting the predictions leading to a thermostabilization with retained enzyme activity, DCA and FoldX performed better and identified 7 and 8 beneficial mutations for Kivd, respectively. HoTMuSiC was very successful in identifying mutations, which led to a less stable or inactive enzyme.

Discussion

Some of the beneficial mutations from the DCA calculations are also calculated to be stabilizing by force field calculations but have a value, is too low to be selected rationally from the dataset. This becomes evident in Figure 3, where the FoldX energies of all individual amino acid substitutions are grouped into classes based on the Gibbs free energy. For example the DCA predicted mutation with the highest experimentally confirmed thermostabilization (V534L) has a FoldX value of -2.6. To identify this mutation in a rational approach based on FoldX alone, more than 200 individual mutations, ranging from FoldX values of -7 to -2, would need to be screened. For the second best mutation M538L (FoldX value +1.2) more than 3500 individual mutations would need to be screened to find this mutation. This

indicates that indeed DCA can find stabilizing mutations, which are hard to identify with other computational methods, suggesting DCA to be a highly complementary method for rational approaches in protein engineering.

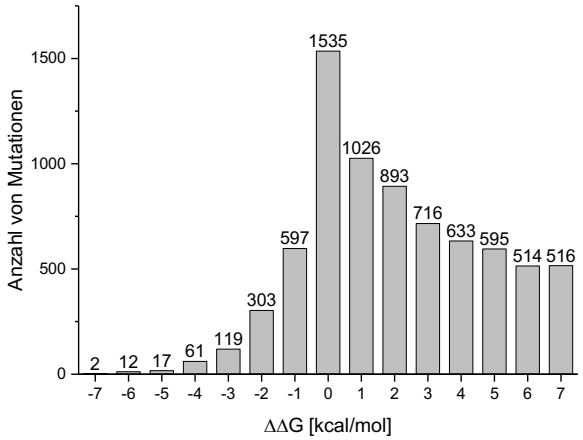

Figure 3: Sums of all amino acid substitutions with similar calculated effects on the Gibbs free energy [kcal/mol] in an exemplary range.

Using Direct coupling analysis (DCA) we could find beneficial mutations that lead to an improved thermostability of the dimeric ketoisovalerate decarboxylase of up to 3.9°C. These mutations were not found using established methods for rational protein design, as recently described (Maier *et al.*, 2018). In context of the 3D structure of the enzyme, the top three mutations with the highest experimentally confirmed increase in T_{50} (S385M, V534L, M538L) are clustered in a region distant from the active site and not exposed to the solvent (Figure 4A). Although no structures of the mutant proteins are available exchanging the amino acids *in silico* can offer limited insight into possible reasons for stabilization. As expected from a structure-based computational approach, the stabilizing effect of the S385M mutation (suggested by the FoldX algorithm) can be explained investigating the structure (Figure 4B, I, II, III). In this case the substitution of serine to methionine reduces the number of hydrogen bonds to neighboring amino acids, however, the positioning of the divalent sulfur atom of the methionine side chain at a distance < 5 Å to the aromatic ring of the phenylalanine side chain at position 381 enables strongly stabilizing hydrophobic interactions which have been described previously to be able to stabilize proteins (Buss *et al.*, 2018; Pal and Chakrabarti, 2001; Tatko and Waters, 2004; Valley *et al.*, 2012). When comparing the wildtype V534 and mutant V534 (Figure 4B, IV and V) the number of hydrogen bonds is unaltered

and no obvious amino acid interactions are disturbed or generated. In the case of the M538L mutation the loss of possible interactions with neighboring phenylalanine residues (Figure 4B, VI) are likely the reason for the FoldX value of +1.2 which indicates a destabilizing mutation. At the same time the newly introduced leucine residue does not generate any clear stabilizing interactions. In both the case of the V534L and the M538L structural reasons for the stabilization are less obvious and thus they were not amongst the top hits in the structure-based predictions.

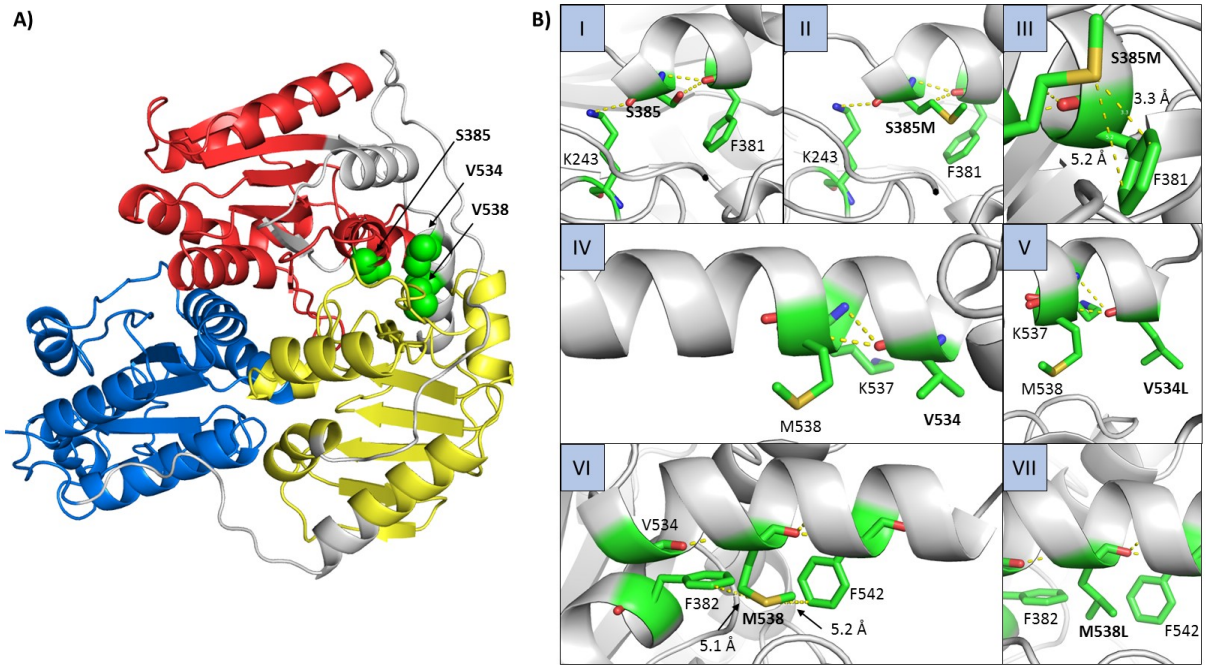

Figure 4: A) Location of the top three experimentally confirmed, stabilizing mutations identified employing DCA and FoldX within the 3D model of the dimeric target enzyme Kivd, based on the KdcA structure (Berthold et al., 2007). B) Detailed structural view on the mutations S385M (I-III), V534L (IV/V) and M538L (VI/VII) showing potential interacting amino acids and hydrogen bonds as identified using the software Pymol.

In summary, we could show that DCA, so far mostly used in structure prediction, could also be used to find beneficial mutations in a protein leading to an improved thermostability. This is especially true, taking into account that for any structure-based prediction of stabilizing mutations high quality 3D structures are necessary and that for rule- or model-based approaches training sets with sufficient coverage of the sequence space are needed, whereas DCA only requires a sufficient number of homologous protein sequences.
Furthermore we conclude, that the interpretation of the DCA energy as an estimate of an evolutionary fitness appears appropriate.

Although our current investigation does not gather sufficient data for true statistical significance, we could show that DCA is a highly promising tool for protein engineers for cases where the structural information of the protein is limited.

Materials and Methods

Protein expression and purification
For heterologous protein expression, *E. coli* BL21(DE3) (E. cloni® EXPRESS BL21(DE3) Electrocompetent Cells, Lucigen) were transformed with the corresponding expression vector according to the manufacturer's instructions and cultivated at 37°C in 30 mL LB medium in a shaking flask overnight. This preculture was then used to inoculate 1 L LB medium (1% (v/v)) for expression culture in a shaking flask. The culture was then incubated at 37°C until the OD_{600} reached a value between 0.6 and 1.0. After lowering the temperature to 25°C, isopropyl-*β*-D-thiogalactopyranosid (IPTG) was added to a final concentration of 0.5 mM and the cells were incubated for an additional 16 hours. The cells were harvested by centrifugation (10000 rcf, 10 min) and resuspended in 30 mL buffer A (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0). Subsequently, the suspension was frozen and stored at -80°C until the start of the purification procedure, but at least for 12 h to aid the cell lysis. In order to start with the purification, the suspension was thawed at 30°C and then incubated with 1 mg/mL lysozyme (AppliChem) and 0.1 mg/mL DNAseI (AppliChem) for 1 h at 37°C in a rotation incubator. After disruption by ultrasonication, the cell lysate was obtained by centrifugation (45000 rcf, 4°C, 1 h), filtered through a 0.45 µm Durapore PVDF membrane (Steriflip, Millipore) and loaded onto a HisTrap FF (1 mL) Ni-NTA column (GE Healthcare, Germany) mounted on an Äkta prime liquid chromatography system. The column was washed with buffer A until the absorption at 280 nm reached baseline-level, followed by changing the buffer to 100% buffer B (50 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole, pH 8.0) and the 6xHis-tagged proteins were eluted. Subsequently the buffer was exchanged using Vivaspin (GE Healthcare) with 1/3 MWCO according to the target protein to PBS-1 (3.5 mM NaH_2PO_4, 8 mM Na_2HPO_4, 50 mM NaCl, pH 7.5) for Kivd, respectively PBS-2 (3.5 mM NaH_2PO_4, 8 mM Na_2HPO_4, 400 mM NaCl, pH 7.5) for the ADH.
The purified proteins were characterized by 16% polyacrylamide gel electrophoresis (PAGE) with standard discontinuous SDS-polyacrylamide Laemmli-midi-gels. The bands were visualized by

staining with 0.01% (w/v) Coomassie Brilliant Blue solution (5 mg Coomassie Brilliant Blue G250, 5 mL H_3PO_4 adjusted with double deionized H_2O to a final volume of 50 mL) and compared to the Page Ruler™ prestained protein ladder (Thermo Fisher Scientific). The concentrations of protein solutions were determined by UV-Vis spectroscopy, using the theoretical molar extinction coefficients at 280 nm, as calculated by the ProtParam tool on the ExPASy Server (Wilkins *et al.*, 1999).

Protein activity assay

In order to determine the T_{50} values, 90 µL (1 µM) of the purified enzymes in buffer C (1 M tris/HCl pH 7.5, 10 mM $MgCl_2$, 10 mM NaCl, 2.5 mM $MgSO_4$, 0.1 mM thiaminepyrophosphate (TPP)) were incubated for 10 min in temperature gradients ranging from 45°C to 64°C. Subsequent to this incubation, the samples were cooled on ice for 5 min. After reequilibration to room temperature, 50 µL of the samples were transferred into a 96-Well plate. 50 µL of the substrate containing test solution (buffer C supplemented with 10 mM ketoisovalerate (KIV), 500 µM NADH, 50 µM TPP, 1 µM ADH) were added to each sample using an automated multi-channel pipette. Subsequently the NADH conversion was monitored via fluorescence measurements (Excitation: 340 nm, Emission: 440 nm).

FoldX and HoTMuSiC *calculations*

Using the FoldX algorithm, ΔΔG values [kcal/mol] were calculated for every possible single amino acid substitution at all amino acid positions throughout the protein as previously described (Maier *et al.*, 2018). Negative ΔΔG values were considered to indicate a stabilizing effect on the enzyme. The HoTMuSiC Server was utilized as described on the respective webpage.

Direct coupling analysis (DCA)

In contrast to conservation analysis, which simply represents the relative abundance of every amino acid at the single positions in the applied database, DCA takes also into account whether amino acids tend to mutate as pairs (coevolution) within the context of their sequence background, as expressed in **Error! Reference source not found.**.

$$\phi_{DCA}(A_1, \ldots, A_L) = -\sum_{i<j} e_{ij}(A_i, A_j) - \sum_i h_i(A_i)$$

Equation 1

(A_i is the amino acid at position i, e_{ij} the (unknown) coupling strength between sites i and j for amino acids A_i and A_j and h_i the local bias term). The model used here is an inverse Potts model (Wu, 1982) derived from statistical physics to describe spins in a crystal lattice. Here, the model is a condensed representation of the alignment in the form of a probability distribution, with the advantage of having decoupled the complex direct dependencies within the alignment from the indirect ones. The parameters of this model are determined in a way that the model can reconstruct the sequence distribution of the alignment. For this purpose a sequence of amino acids $A_1, \ldots, A_L$ is assigned the probability P, shown in **Error! Reference source not found.**.

$$P(A_1, \ldots, A_L) = \frac{1}{Z} \exp\left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right)$$

Equation 2

The associated partitioning function Z (**Error! Reference source not found.**) represents the summation of all possible sequences so that the closing condition $\Sigma\ P = 1$ applies. The model parameters e_{ij} and h_i must be determined by the alignment. In order to determine the coupling parameters X computationally efficiently, there are different approaches. In the implementation used here, the mean field approximation was used, which averages the effect of all residues on a single residue instead of recalculating the effect of the remaining residues for each individual residue (mfDCA) Specifically, this means that the coupling matrix can be represented according to **Error! Reference source not found.** with the covariance matrix shown in **Error! Reference source not found.**. Details have been described elsewhere (Morcos *et al.*, 2011) and the interested reader can find more information in a recent review (Zerihun and Schug, 2017).

$$e_{ij}(A_i, A_j) = -C_{ij}^{-1}(A_i, A_j)$$

Equation 3

$$c_{ij} = -P_{ij}(A_i, A_j) - P_i(A_i)P_j(A_j)$$

Equation 4

Exponent ϕ_{DCA} (**Error! Reference source not found.**) can be interpreted as energy by analogy to other applications of the Potts model in statistical physics. To estimate the effect of mutation A_i->B the difference of ϕ upon mutation (equation 5) is calculated (Figliuzzi *et al.*, 2016).

$$\Delta\phi(A_i \rightarrow B) = \phi_{mut} - \phi_0$$

Equation 5

Acknowledgements

K.S.R and M.M. acknowledge funding via the Helmholtz programme "BioInterfaces in Technology and Medicine". J.E. and A.S. are supported by the Helmholtz Association Initiative and Networking Fund under project number ZT-I-0003. We thank Anke Dech for help with the protein purification.

References

Abdel-Banat, B. M. A. *et al.* (2010). High-temperature fermentation: how can processes for ethanol production at high temperatures become superior to the traditional process using mesophilic yeast? Appl Microbiol Biot. *85*, 861-867.

Berthold, C. L. *et al.* (2007). Structure of the branched-chain keto acid decarboxylase (KdcA) from Lactococcus lactis provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction. Acta Crystallogr D Biol Crystallogr. *63*, 1217-1224.

Bornscheuer, U. T. *et al.* (2012). Engineering the third wave of biocatalysis. Nature. *485*, 185-194.

Broom, A. *et al.* (2017). Computational tools help improve protein stability but with a solubility tradeoff. J Biol Chem. *292*, 14349-14361.

Buss, O. *et al.* (2018). Improvement in the Thermostability of a beta-Amino Acid Converting omega-Transaminase by Using FoldX. Chembiochem. *19*, 379-387.

Dago, A. E. *et al.* (2012). Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proc Natl Acad Sci U S A. *109*, E1733-1742.

Das, R. and Gerstein, M. (2004). A method using active-site sequence conservation to find functional shifts in protein families: Application to the enzymes of central metabolism, leading to the identification of an anomalous isocitrate dehydrogenase in pathogens. Proteins. *55*, 455-463.

El-Gebali, S. *et al.* (2018). The Pfam protein families database in 2019. Nucleic Acids Res.

Figliuzzi, M. *et al.* (2016). Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol Biol Evol. *33*, 268-280.

Figliuzzi, M. *et al.* (2016). Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol Biol Evol. *33*, 268-280.
Finn, R. D. *et al.* (2015). HMMER web server: 2015 update. Nucleic Acids Res. *43*, W30-38.
Kristjansson, J. K. (1989). Thermophilic Organisms as Sources of Thermostable Enzymes. Trends Biotechnol. *7*, 349-353.
Kyte, J. and Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. J Mol Biol. *157*, 105-132.
Lin, P. P. *et al.* (2014). Isobutanol production at elevated temperatures in thermophilic Geobacillus thermoglucosidasius. Metab Eng. *24*, 1-8.
Maier, M. *et al.* (2018). On-Demand Production of Flow-Reactor Cartridges by 3D Printing of Thermostable Enzymes. Angew Chem Int Ed Engl. *57*, 5539-5543.
Modarres, H. P. *et al.* (2016). Protein thermostability engineering. Rsc Adv. *6*, 115252-115270.
Morcos, F. *et al.* (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. P Natl Acad Sci USA. *108*, E1293-E1301.
Pal, D. and Chakrabarti, P. (2001). Non-hydrogen bond interactions involving the methionine sulfur atom. J Biomol Struct Dyn. *19*, 115-128.
Potapov, V. *et al.* (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel. *22*, 553-560.
Pucci, F. *et al.* (2016). Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. Sci Rep. *6*, 23257.
Romero, P. A. and Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol. *10*, 866-876.
Schug, A. *et al.* (2009). High-resolution protein complexes from integrating genomic information with molecular simulation. Proc Natl Acad Sci U S A. *106*, 22124-22129.
Schug, A. *et al.* (2009). High-resolution protein complexes from integrating genomic information with molecular simulation. Proc. Natl. Acad. Sci. U. S. A. *106*, 22124-22129.
Schwab, T. and Sterner, R. (2011). Stabilization of a metabolic enzyme by library selection in Thermus thermophilus. Chembiochem. *12*, 1581-1588.
Schymkowitz, J. *et al.* (2005). The FoldX web server: an online force field. Nucleic Acids Res. *33*, W382-W388.
Soh, L. M. J. *et al.* (2017). Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design. ACS Synth Biol. *6*, 610-618.
Song, Y. *et al.* (2013). High-resolution comparative modeling with RosettaCM. Structure. *21*, 1735-1742.
Sutiono, S. *et al.* (2018). Structure-Guided Engineering of alpha-Keto Acid Decarboxylase for the Production of Higher Alcohols at Elevated Temperature. ChemSusChem. *11*, 3335-3344.
Tatko, C. D. and Waters, M. L. (2004). Investigation of the nature of the methionine-pi interaction in beta-hairpin peptide model systems. Protein Sci. *13*, 2515-2522.
Turner, P. *et al.* (2007). Potential and utilization of thermophiles and thermostable enzymes in biorefining. Microb Cell Fact. *6*, 9.
Uguzzoni, G. *et al.* (2017). Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. Proc Natl Acad Sci U S A. *114*, E2662-E2671.
Valley, C. C. *et al.* (2012). The methionine-aromatic motif plays a unique role in stabilizing protein structure. J Biol Chem. *287*, 34979-34991.
Wang, C. *et al.* (2012). Improving the thermostability of alpha-amylase by combinatorial coevolving-site saturation mutagenesis. BMC Bioinformatics. *13*, 263.
Weigt, M. *et al.* (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A. *106*, 67-72.
Wilkins, M. R. *et al.* (1999). Protein identification and analysis tools in the ExPASy server. Methods Mol Biol. *112*, 531-552.
Wu, F. Y. (1982). The Potts-Model. Rev Mod Phys. *54*, 235-268.

Xiao, H. *et al.* (2015). High Throughput Screening and Selection Methods for Directed Enzyme Evolution. Ind Eng Chem Res. *54*, 4011-4020.
Zamost, B. L. *et al.* (1991). Thermostable Enzymes for Industrial Applications. J Ind Microbiol. *8*, 71-81.
Zeiske, T. *et al.* (2016). Thermostability of Enzymes from Molecular Dynamics Simulations. J Chem Theory Comput. *12*, 2489-2492.
Zerihun, M. and Schug, A. (2017). Biomolecular coevolution and its applications- going from structure prediction towards signaling, epistasis, and function. Biochem. Soc. Transact. *45*, 1253-1261.

Table and figure legends