Received: 12 April 2019 | Revised: 2 October 2019 | Accepted: 21 October 2019

DOI: 10.1096/fj.201900948RR

RESEARCH ARTICLE

THE FASEB JOURNAL

Algorithm-based coevolution network identification reveals key functional residues of the α/β hydrolase subfamilies

Zhiyun Wu[1] | **Hao Liu**[1] | **Lishi Xu**[1] | **Hai-Feng Chen**[1,2] | **Yan Feng**[1]

[1]State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

[2]Shanghai Center for Bioinformation Technology, Shanghai, China

Correspondence
Hai-Feng Chen and Yan Feng, State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.
Email: haifengchen@sjtu.edu.cn (H-F. C.); yfeng2009@sjtu.edu.cn (Y. F.)

Funding information
National Natural Science Foundation of China (NSFC), Grant/Award Number: 31770771 and 31620103901; Research and Development, Grant/Award Number: 2017YFE0103300, 31770771, 31620103901 and YG2017MS08; Shanghai Jiao Tong University

Abstract

Covariant residues identified by computational algorithms have provided new insights into enzyme evolutionary routes. However, the reliability and accuracy of routine statistical coupling analysis (SCA) are unable to satisfy the needs of protein engineering because SCA depends only on sequence information. Here, we set up a new SCA algorithm, SCA.SIM, by integrating structure information and MD simulation data. The more reliable covariant residues with high-quality scores are obtained from sequence alignment weighted by residual movement for eight related subfamilies, belonging to α/β hydrolase family, with *Candida antarctica* lipase B (CALB). The 38 predicted covariant residues are tested for function by high-throughput quantitative evaluation in combination with activity and thermostability assays of a mutant library and deep sequencing. Based on the landscapes of both activity and thermostability, most mutants play key roles in catalysis, and some mutants gain 2.4- to 6-fold increase in half-life at 50°C and 9- to 12-fold improvement in catalytic efficiency. The activity of double mutants for A225F/T103A is higher than those of A225F and T103A which means that SCA.SIM method might be useful for identifying the allosteric coupling. The SCA.SIM algorithm can be used for protein coevolution and enzyme engineering research.

KEYWORDS

algorithm optimization, enzyme, mutant library, protein engineering, statistical coupling analysis

1 | INTRODUCTION

Given the rapid growth in available genome data from many organisms, it has become possible to apply statistical sequence analysis to determine the relationships among the amino acid sequence of a protein, its function, and the three-dimensional structure. A set of methods has been used to reveal the groups of residues that are jointly involved in determining functional properties.[1-4] Statistical coupling analysis (SCA) performs covariation analysis of a large number of multiple sequences and is capable of identifying sets of residues that are important for protein folding,[5] allostery,[6] enzymatic activity,[7] and thermal stability.[8] Recently, Salinas and

Abbreviations: CALB, *Candida antarctica* lipase B; HPC, high-performance computing; MD, molecular dynamics; MSAs, multiple sequence alignments; NVT, network virtual terminal; *p*NP, *p*-nitrophenol; PME, particle mesh Ewald; PBS, sodium phosphate buffer; RTA, real-time analysis; SCA, statistical coupling analysis; SCA.SIM, statistical coupling analysis by integrating structure information and MD simulation data.

Ranganathan using SCA successfully capture the energetic couplings that contribute to protein function which proves the reliability of this method.[9] Covariant residues are distributed around protein active sites but can connect to distant functional surface sites through pathways of residue interactions throughout the protein.[10,11]

The combination between amino acid and codon information is used to improve the contact prediction.[12] Also, the SCA method combined with MD simulations (SCA·MD) reveals that protein dynamics plays a key role in enzyme catalysis.[13] Despite their improved accuracy, the MD simulation-assisted SCA methods cannot overcome all the limitations of the sequence-based SCA methods, suggesting a need for further development. Because the structure is more conservation than sequence, structure-based process could construct high-quality alignment. Therefore, a more accurate SCA method that accounts for sequence, structure, and dynamics is an important need.

Here, we set up a new SCA algorithm named SCA.SIM to obtain key functional covariant residues defined as groups of statistically correlated amino acid positions in a protein family identified by their combined contributions to the most significant eigenmodes of a preconstructed SCA correlation matrix. The algorithm is based on the structure alignment information from the 3DM database[14] and the dynamics information from molecular dynamics (MD) simulation. Specifically, a defined structural feature was used to guide the classification of the enzyme subfamilies for increasing the evolutionary structural homology and decreasing the historical noise, then a residue-level dynamical correlation matrix based on MD simulation was incorporated into a new SCA matrix for SCA.SIM (Figure 1). To evaluate this approach, 38 covariant positions were predicted for *Candida antarctica* lipase B (CALB) and confirmed with function annotation and site-saturated mutagenesis assay. The results indicate that this approach might have great potential for the coevolution analysis of proteins and useful tool for the directed evolution design of enzymes.

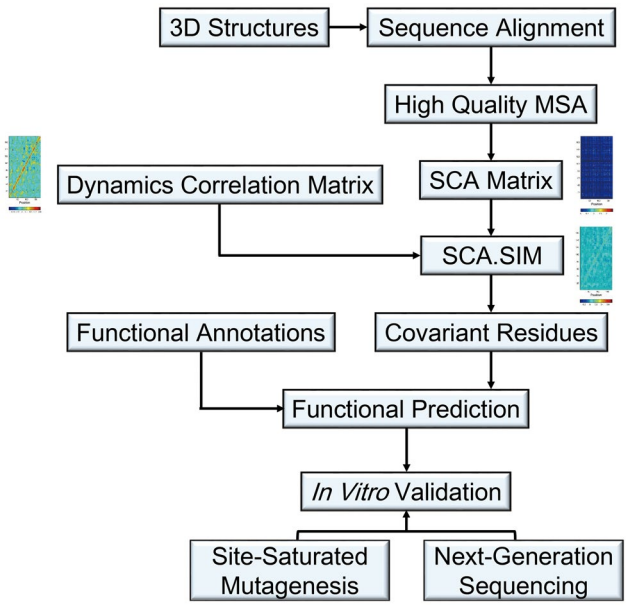

FIGURE 1 Flowchart of the SCA.SIM system

2 | MATERIALS AND METHODS

2.1 | Structure-based sequence alignment with 3DM

The α/β hydrolase (2014) 3DM database is a high-quality structure-based multiple sequence alignment based on 81624 sequences and composed of separate subfamily sequence alignments of 204 subfamilies for which a structure is available.[14] A phylogeny tree analysis of the representative structures of 204 subfamilies was conducted and is shown in Figure S1. For a more specific analysis, we classified the subfamilies according to the cap structure and their phylogenic relationships.[15] CALB belongs to Cluster 1 and consists of eight subfamilies with a total of 2182 sequences. We used these sequences to generate a subset of the α/β hydrolase (2014) 3DM database, using the same alignment algorithm within the subset. Then, we obtained a high-quality structure-based multiple sequence alignment of families closely related to CALB. The alignment consisted of 2182 sequences and had 177 aligned positions.

2.2 | Molecular dynamics simulation

The CALB protein model was extracted from the Protein Data Bank (pdb code: 1TCA).[16] The docking of the protein and p-nitro phenyl caprylate C8 was conducted by Maestro9.2.[17] The simulation was conducted using the AMBER14 software package.[18] Counterions were used to maintain system neutrality. The system was solvated in a truncated octahedron box of TIP3P waters with a buffer of 10 Å. The particle mesh Ewald (PME) method[19] was employed to treat long-range electrostatic interactions. The *ff12SB* force field was used for the protein. Antechamber was used to generate the force field of the substrate.[20] Bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[21] A 22500-step steepest descent minimization was performed, followed by heating and brief equilibration for 20 ps in the NVT ensemble at 321 K. Three independent trajectories of 50 ns each were simulated.

2.3 | Movement correlation matrix

Correlations between all pairs of residues in the CALB complex were calculated from the covariance matrix of equation (1) [22-25]:

$$C_{ij} = \frac{\Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t)}{\sqrt{\Delta\vec{r}_i(t)^2 \, \Delta\vec{r}_j(t)^2}} \quad (1)$$

where $\Delta\vec{r}_i(t) = \vec{r}_i(t) - \vec{r}_i(t)$, $\langle\cdot\rangle$ is the time averaging and $\vec{r}_i(t)$ is the position of node i at time point t. In this study, we constructed correlation-based networks[26] using the covariance matrices along the last 50 ns in each trajectory, with 40 ps per snapshot. An edge is defined between any two nodes without covalent bond but with heavy atoms closer than 4.5 Å over 50% sampling time. The network topological parameters were calculated with Cytoscape 3.1.1.[27] The shortest path between any two nodes was calculated using the Floyd-Warshall algorithm.[28] For community analyses, the Girvan-Newman algorithm[29] was utilized with the network tools developed by the Luthey-Schulten group.[26,29]

2.4 | SCA and SCA.SIM

The coevolution relationships between different positions in MSA were explored by SCA calculation.[8] The input for SCA calculation is the multiple sequence alignment, and the output is a four-dimensional matrix whose size is L × L × 20 × 20 for an MSA that has L positions and 20 amino acids in its sequence. We can generate a two-dimensional position correlation matrix named SCA after reducing the dimensionality of the SCA matrix. The SCA.SIM matrix was created by multiplying the individual elements of SCA by the corresponding elements of the truncated MD movement correlation matrix.[13] In this study, instead of fully applying the SCA.MD method from R. August Estabrook's work,[13] we only used movement correlation to weight the SCA matrix. Then, we used the SCA.SIM matrix instead of the SCA matrix as the input to calculate the covariant residues, which was described in the literature.[8] To evaluate the statistical significance of difference between SCA and SCA.SIM, the Wilcoxon signed-rank test was done with the data in Table S1 using R 3.5.0 software. The command Wilcox.test embedded in R was used to calculate the *P* value.

2.5 | Library creation and screening

The CalB gene was synthesized at GenScript Crop. The mutants were prepared by whole-plasmid PCR with primers containing NNS (sense strand)/SNN (antisense strand) degeneracy at each of the 38 covariant sites. All the primers used in this study were synthesized from Invitrogen (Shanghai, China) and are shown in Table S2 in the supplemental material. PCR was performed with Proofast™ (ATG Biotechnology Co., Ltd) Super Fidelity polymerase and a temperature program consisting of 95°C for 2 min; 30 cycles of 10 s at 95°C, 15 s at 55°C, and 7 min at 72°C; and a final 10 min extension at 72°C. The modified PCR program was used for the construction of paired-residue libraries (A225/T103). The libraries were electroporated into Rosetta (DE3) cells. The clones were picked in 96-well plates containing 200 μL of two YT medium with 100 μg mL^{-1} ampicillin. After growth at 37°C for 12 h, the cultures were used to inoculate fresh medium in an identical plate and incubated for a further 3 h at 37°C. After 0.1 mM isopropyl 1-thio-β-D-galactopyranoside was added, the cells were incubated for 24 h at 15°C.

Cells were harvested by spinning at 4000 rpm for 30 min and lysed by a triple freeze-thaw from −80°C. Then, 200 μL of 50 mM sodium phosphate buffer (PBS, pH 7.5) was added to each well. The supernatant was divided into two 96-well PCR plates. One plate was incubated at high temperature for 15 min and cooled at 4°C for 20 min. The other plate was incubated under identical conditions except for the incubation at high temperature. The ratio of the activity of each clone was taken as the residual activity, which was used to identify the positive clones.

2.6 | Enzyme activity assays

*p*NP caprylate was used to compare the activity of CalB variants. The ability of the enzyme to hydrolyze *p*NP caprylate was determined by measuring the absorbance of *p*-nitrophenol liberated using a UV-2550 spectrophotometer with a thermal control unit (Shimadzu, Kyoto, Japan). The reaction mixture consisted of 0.02 mL of 10 mM *p*NP caprylate as a substrate in acetonitrile and 0.97 ml of 50 mM PBS, pH 7.5 containing an appropriate amount (10 μL) of the enzyme solution. The enzyme reaction was performed for 1 min at 37°C. The activities were determined photometrically at 37°C, and the buffer was adjusted at 37°C unless otherwise stated. One lipase unit in this assay is defined as the amount of enzyme that liberates 1 μ moL of *p*-nitrophenol/min under these conditions.

For the kinetic studies, the concentration of *p*NP caprylate increased from 1.5 to 500 μM. The enzymatic activity of CalB variants was determined at 37°C. Kinetic parameters V_{max} and K_m were acquired by fitting enzymatic activities as a function of substrate concentrations to the Michaelis-Menten equation using non-linear regression of the software Origin8.0. The parameter k_{cat} was obtained using the following equation: $k_{cat} = V_{max}/[E]$, where [E] is the molar concentration of the enzymes.

2.7 | Thermal inactivation and unfolding

The CalB variants (0.1 mg/mL) were incubated at different temperatures for different time intervals from 0 to 150 min

and then cooled on ice for 10 min. Their residual enzyme activities were assayed at 37°C as described above. The data were fitted to first-order plots and analyzed with the first-order rate constants (k_d) determined by linear regression of ln (residual activity) *versus* the incubation time (t). The time required for the residual activity to be reduced to half ($t_{1/2}$) of the CalB variants was calculated using the following equation: $t_{1/2} = \ln 2/k_d$. The changes in transition state free energy ($\Delta\Delta G$) for inactivation between mutants and wild type were calculated using the following equation: $\Delta\Delta G = -RT \ln (k_d \text{ mutant}/k_d \text{ wild type})$[30] where T and R are temperature and the gas constant (1.987 $\text{cal}\cdot\text{K}^{-1}\cdot\text{mol}^{-1}$), respectively. $\Delta\Delta G$ of wild type was used as a reference value. k_d wild type and k_d mutant were inactivation rate constants of wild-type and mutant CalB, respectively.

2.8 | Solexa sequencing

According to the activity and thermostability data, we collected clones with greater activity or thermostability than wild type as the sorted cell populations. All clones were defined as unsorted cell populations. Then, we had three libraries: activity-sorted cell populations, thermostability-sorted cell populations, and unsorted cell populations. All samples were processed in parallel and sequenced in same run to minimize potential biases. These three libraries were diluted into LB plus ampicillin and grown for 12 h at 37°C with shaking (250 r.p.m.), respectively. Overnight cultures were centrifuged and miniprepped. Purified DNA was quantified, and 100 ng of plasmid DNA per 50 μL PCR reaction was used as a template to obtain targeted fragments by PCR amplification. The PCR products were purified and eluted in 20 μL of dH_2O. Purified PCR products were quantified in triplicate using lambda-DNA as a standard. All libraries were pooled in equimolar proportions and sequenced using a MiSeq, version 3, 2 × 300 run by the GENEWIZ Company.

Sequences from the Illumina RTA base caller were imported into CLC Genomics Workbench as ".qseq" files and trimmed for quality using a cut-off of 0.05 for the modified Mott algorithm. Bases that did not pass the trim filter were deleted from each read, and reads shorter than 49 bp were discarded. Custom software written in MATLAB was used to count the number of occurrences of each allele in each population. The functional effect of each allele was calculated from the frequency of each amino acid at each position.

3 | RESULTS

The protocol of the SCA.SIM system includes structure-based multiple sequence alignment and is shown in Figure 1. These high-quality multiple sequence alignments (MSAs) will be used as inputs to construct the SCA matrix. At the same time, multiple MD simulation trajectories will represent the representative structure and will be used to calculate the dynamic correlation matrix. A combination of these matrices will create the SCA.SIM system. Functional annotations and in vitro experiments are used to validate this system.

3.1 | Construction of SCA.SIM matrix

Lipases (EC 3.1.1.3) are highly abundant across species and have a rich evolutionary history; therefore, this class of enzymes was mainly used to evaluate the SCA algorithm. Among the characterized lipases, *Candida antarctica* lipase B (CALB) has been extensively studied, including biochemical and site-directed mutagenesis of the active site residues involved in its stereoselectivity and activity.[31,32] These findings make CALB an excellent model for analyzing the function of covariant residues in the α/β hydrolases subfamilies. Because the function of an enzyme is correlated with its dynamic motion, combining SCA and MD correlations identifies the group of residues that are correlated both evolutionarily and in structural motion. MD and SCA are independent covariation analyses that have several similarities: they reveal couplings between any amino acid and other amino acids within a given protein, produce matrices with pronounced diagonal elements, and relate peptide regions that can be spatially distant.[13] It is mathematically straightforward to combine the two sets of correlations by multiplying the two matrix elements, resulting in a dynamics-weighted SCA matrix (see Materials and Methods) (Figure 2).

3.2 | SCA.SIM reveals more compact covariant residues than SCA

The SCA.SIM matrix obtained for CALB shows a similar pattern to those of SCA and the motion correlation matrices (Figures 3A-D). Significantly, the residue positions in SCA.SIM are more compact than those in SCA, resulting in a more defined section. The first mode describes a "coherent" correlation between all positions, as reported in the literature.[8] Historical noise which was caused by phylogenetic bias and independent functional constraints and did not include evolutionary correction, while evolutionary correlations and common ancestry represent evolutionary conservation and should have strong relation with biological function, is expected to exclude coherent correlations between sequence positions. This phenomenon is why a pointed distribution appears starting from the origin on

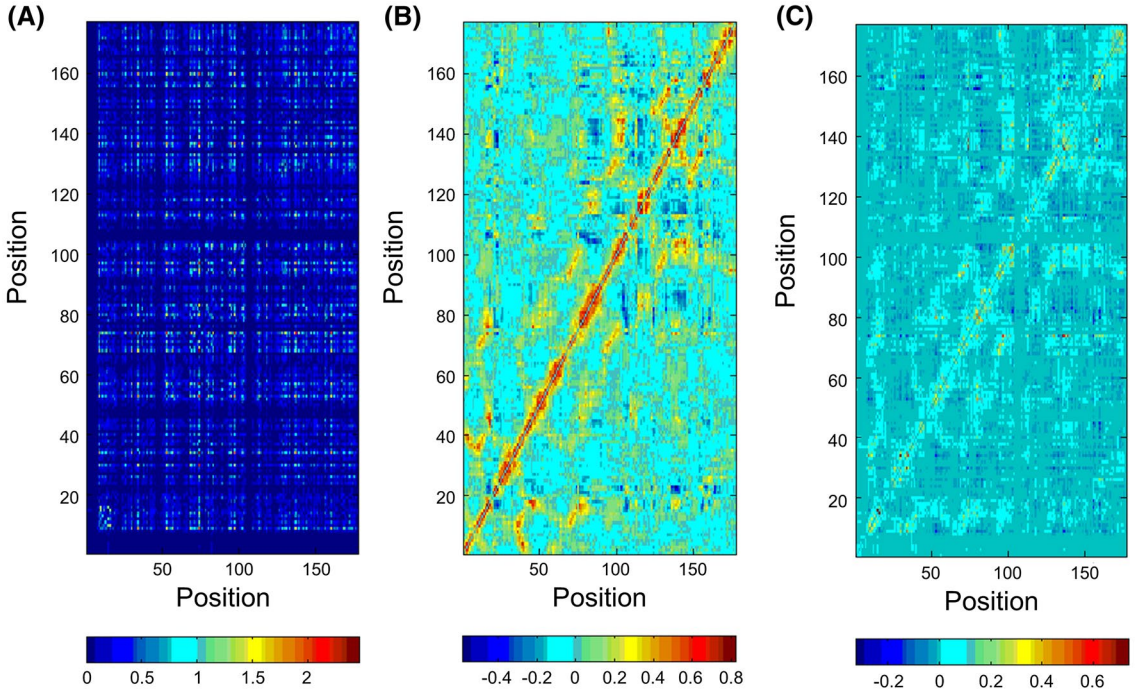

FIGURE 2 SCA.SIM analysis. For better presentation of all the matrices, the values on the diagonal were set to 0. A, The position correlation matrix obtained by reducing the dimension of the SCA matrix of CALB-related families (SCA). The correlations vary from 0 (uncorrelated) to 2.9 (correlated). Some columns and rows are missing due to deletions in the multiple sequence alignment. B, The truncated movement correlation of a 50-ns simulation of the CALB-substrate complex system. The correlations vary from −0.48 (anti-correlated) to 0.8 (correlated). C, The SCA. SIM matrix, generated by multiplying the individual elements of SCA with corresponding elements of the truncated movement correlation matrix, varying from −0.26 (anti-correlated) to 0.71 (correlated)

eigenvector 1. After we weight the SCA with movement correlation, these "coherent" distributions are corrected. That is, the historical noise is weakened after the SCA is weighted. This correction ensures that the statistical correlations are contributed mostly by evolutionary correlations. Meanwhile, this correction distinguishes between correlations that reflect physical interactions and other sources such as common ancestry.

The positional distribution analysis of the top three eigenvectors is shown in Figure 4A-B. We analyze the probability distribution of the eigenmode values of the first eigenvector. All the distributions are fitted with a T distribution, and the covariant residues are determined by taking the 10% positions from the fitted distribution. The eigenmode values are more concentrated around the origin in SCA.SIM, while the difference between the maximum and minimum values is larger than that in the original SCA. In the original SCA, the distribution of eigenmode values is relatively uniform, and the significant positions may have a similar eigenmode to those of insignificant positions. This situation will easily cause false selection when we want to select only significant positions by taking a certain proportion of the fitted distribution. However, in SCA.SIM, the difference between significant and insignificant positions is enlarged, which improves the precision in selecting only significant positions.

We find a total of 38 covariant residues by SCA.SIM and 27 by the original SCA. The covariant residues are then projected onto the CALB protein structure (Figure 4C). In contrast to the covariant residues identified by the original SCA, the covariant residues identified by SCA.SIM are more connected and distributed more tightly around the binding pocket.

3.3 | In silico function validation

We identified 63 residues in CALB that are annotated as functional according to the literature of mutagenesis analyses.[31-49] At first, we directly statistics the overlap functional residues predicted by SCA and SCA.SIM and the results are shown in Figure 5A. This indicates that SCA.SIM performs better than SCA. To evaluate the significance of difference, the Wilcoxon signed-rank test was used to calculate the *P* value, in which the median line is symmetrical and the data meet the assumption of test (shown in Figure 5B). The *P* value is 1.07×10^{-5} and much smaller than 0.05. This suggests that the SCA.SIM has more significant prediction ability of functional residues than SCA.

To further evaluate the function of unannotated covariant residues, the correlation/anti-correlation was calculated

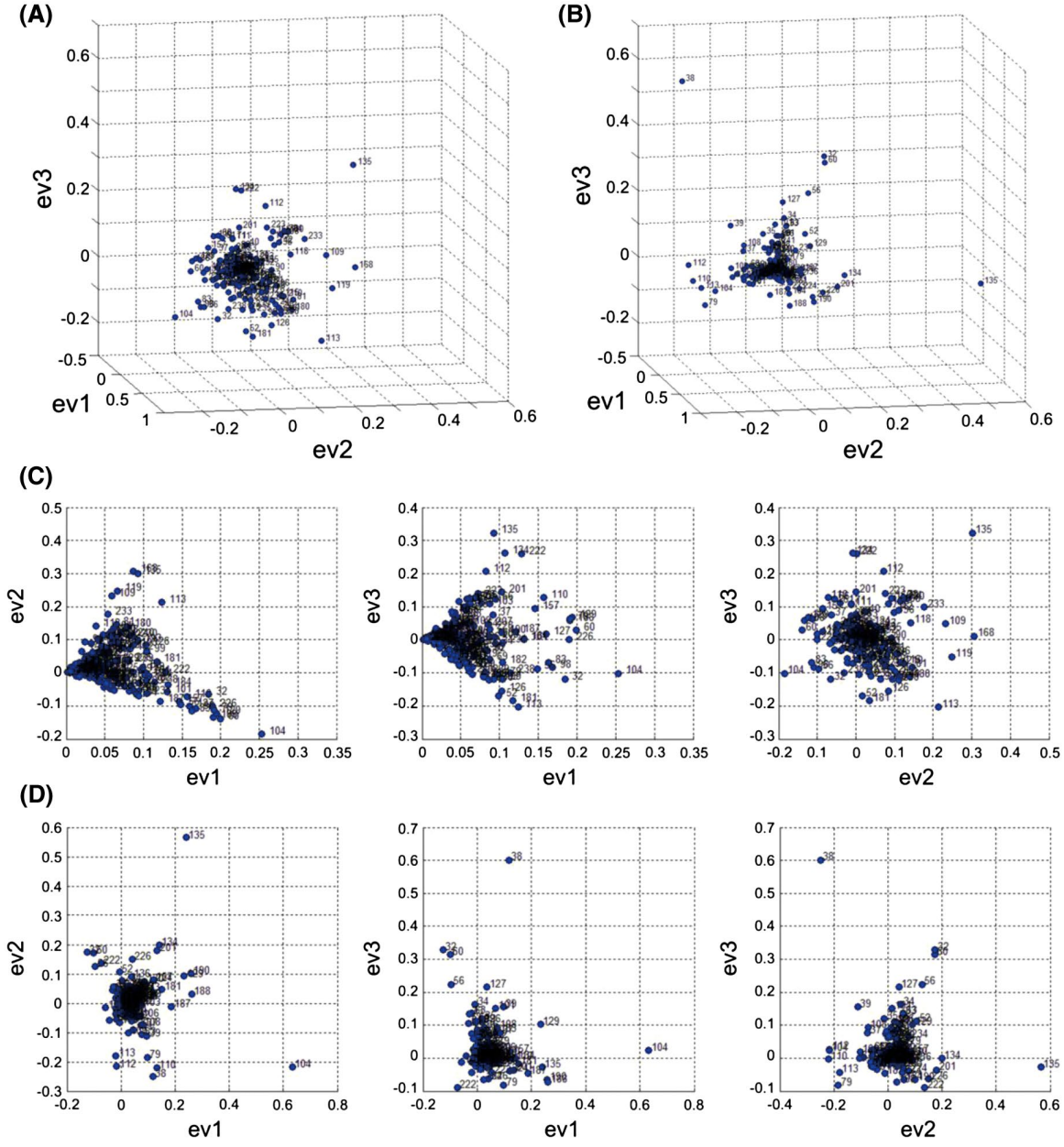

FIGURE 3 Comparison of covariant residues generated by the SCA matrix and the SCA.SIM matrix. A, The 3D positional distribution of the top three eigenvectors of the SCA matrix. B, The 3D positional distribution of the top three eigenvectors of the SCA.SIM matrix. C, The 2D positional distribution of the top three eigenvectors of SCA. D, The 2D positional distribution of the top three eigenvectors of SCA.SIM

between the covariant residues and annotated functional residues based on the previous method.[13] Top 20 ranked pairs of covariant residue were defined as predicted functional ones and their spatial distribution in CALB (shown in Figure S2). The predicted functional covariant residues are shown in Figure 6A whose function includes the activity, thermostability, and enantioselectivity. And the residues in red are reported functional sites.[31,32,38,41,42] For example, E188 has a strong correlation with W104, D134, and M129 (Figure 6B). An increase in catalytic activity was observed after carrying out a single point mutation on these three sites. In addition, E188 is correlated with V190, and the two are connected by I189 (Figure 6C), which is reported to be involved in catalytic activity and enantioselectivity.[31] Accordingly, we speculate that E188 is related to both enzyme activity and enantioselectivity. Y135 exhibits significant correlations with D134, V190, D187, H224, W104, and M129 (Figure 6D). Mutations of D134, W104, and M129 can increase the enzyme catalytic activity, whereas mutations of V190 can decrease the catalytic activity. Moreover, D187 and H224 belong to the catalytic triad. Therefore, Y135 is very likely to affect the catalytic

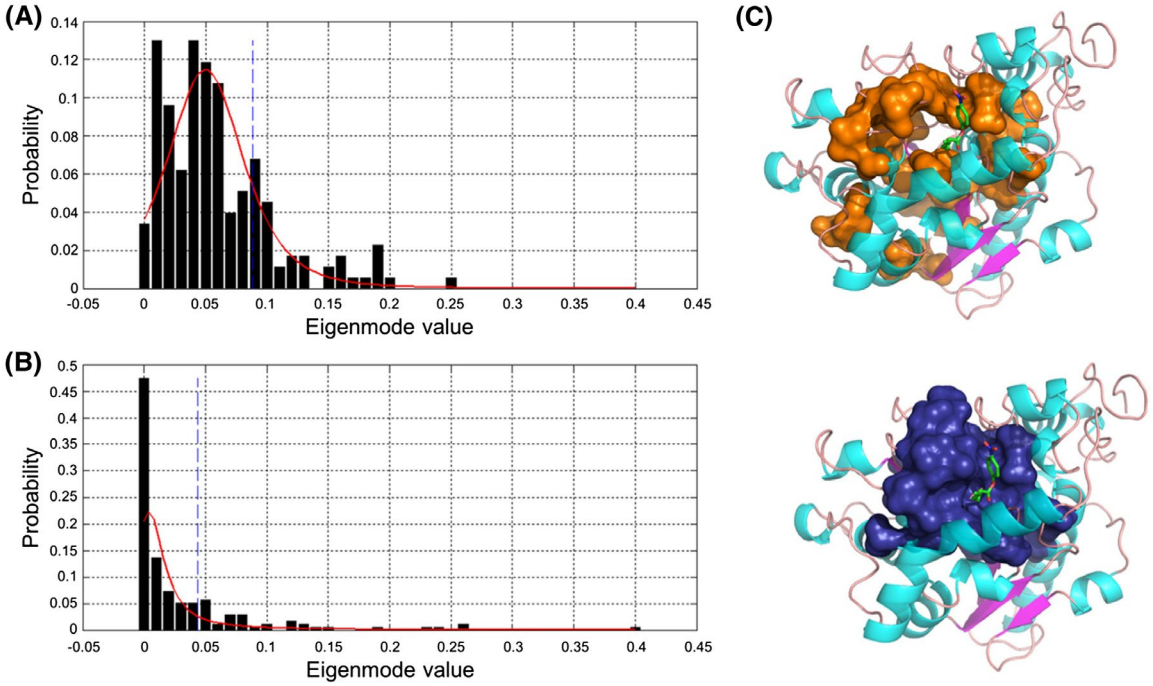

FIGURE 4 Probability distribution analysis of first eigenmode values and covariant residues on tertiary structure of CALB. A, The probability distribution analysis of the first eigenmode values of the original SCA. B, The probability distribution analysis of the first eigenmode values of the SCA.SIM. C, Covariant residues obtained for the tertiary structure of CALB: covariant residues (yellow) generated with the original SCA, covariant residues (blue) generated with SCA.SIM

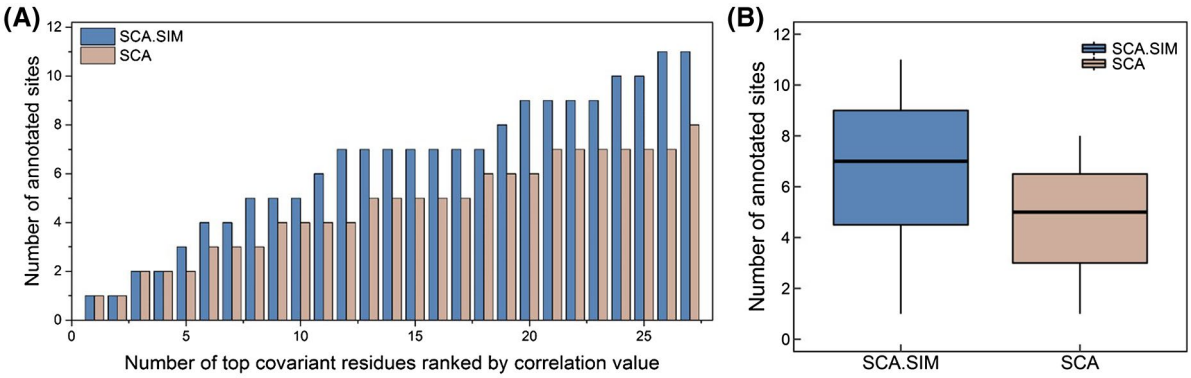

FIGURE 5 Comparison of known functional residues among the covariant residues obtained by SCA and SCA.SIM. A, The number of known functional residues among the covariant residues obtained by both methods. B, Data distribution of known functional residues. Due to the two methods generate different numbers of covariant residues, most of the covered known functional residues are gathered in the high eigenmode value region. Therefore, we chose the top 27 ranked residues for the presentation. The x-axis represents the accumulated number of covariant residues which are ranked by eigenmode values. The y-axis represents the number of known functional residues corresponding to these covariant residues

activity. N181 is spatially related to multiple thermostability-related residues and, in addition, exhibits significant correlations with W104 and M129 (Figure 6E). Therefore, N181 is likely to be involved in catalytic activity and thermostability. However, the function of five sites for 37, 182, 207, 132, and 179 are not identified from the literature. Therefore, in vitro experiment was further used to verify their function.

3.4 | In vitro experimental verification

To further evaluate the function of the covariant residues, we used a quantitative high-throughput method based on next-generation sequencing for the large-scale mutational analysis of proteins in a cellular context. The method involves three steps:[1] construct saturated mutant libraries of 38 covariant sites and measure the activity and thermostability

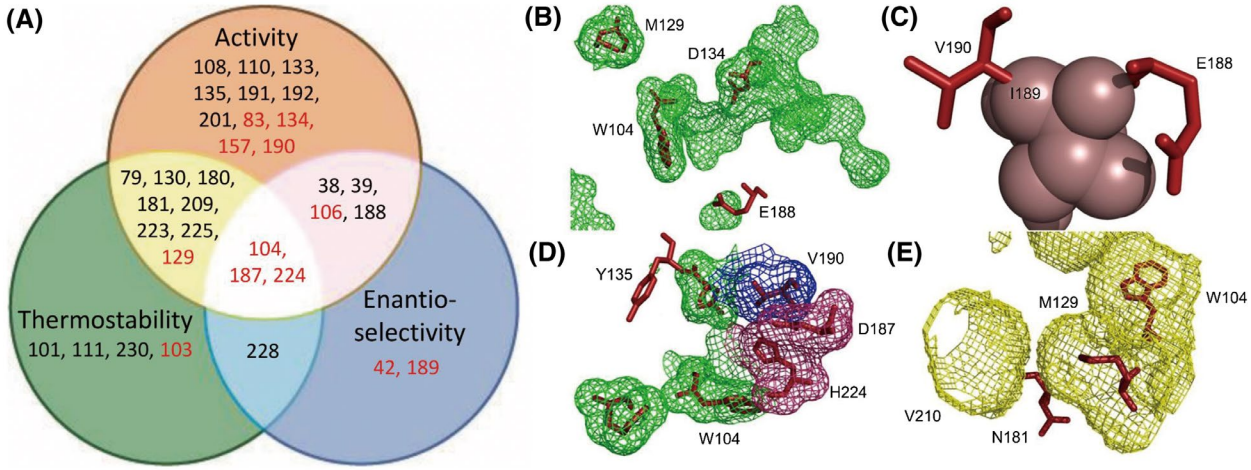

FIGURE 6 Function prediction of covariant residues with top ranked eigenmode value and biological properties. Residues shown in red stick representation have significant correlations with each other. A, Function prediction of covariant residues, including activity, thermostability, and enantioselectivity. The residues in red are reported functional sites. B, Correlations between covariant residue E188 and catalytic activity-related residues (green mesh). C, Enantioselectivity-related residue I189 is shown in brown sphere representation. D, Correlations between covariant residue Y135 and known functional residues. Residues shown in green mesh are reported to increase catalytic activity after mutation. Resies shown in blue mesh are reported to decrease catalytic activity after mutation. Residues shown in red mesh are catalytic residues. E, Evolution and spatial correlations between covariant residue N181 and thermostability-related residues (yellow mesh)

of all the libraries;[2] perform a cell selection step, in which clones with activity or thermostability above a specified threshold are selected (Figure S3); and [3] perform Solexa high-throughput sequencing to determine the frequency of each allele in the unselected and selected clones.[50-52] The effect of each mutation is then expressed as the log frequency of observing each amino acid x at each position i in the selected (sel) versus the unselected (unsel) population relative to the wild type (WT) using equation (2):

$$\Delta E_i^x = \log\left[\frac{f_i^{x,sel}}{f_i^{x,unsel}}\right] - \log\left[\frac{f_i^{WT,sel}}{f_i^{WT,unsel}}\right] \quad (2)$$

In this assay, mutations without functional effects should show a similar frequency in the selected population and the wild type ($\Delta E_i^x \approx 0$); otherwise, mutations should provide a quantitative measure of the functional effect. We used this method to carry out a complete single mutation scan of the 38 covariant sites, which were individually mutated to every other amino acid (Figure 7, 38 positions × 19 mutations + wild type = 723 variants).

In general, most mutations show significant sensitivity to activity or thermostability, which indicates that covariant positions are functionally important. To quantitatively evaluate the performance of SCA.SIM, the average (AVG) values shown in Figure 7 are used to analyze the activity or thermostability effect of each mutation site. The positive or negative value indicates the gain or loss of protein functions. Thus, the absolute AVG values can be used to indicate the degree of functional changes. The average of the absolute AVG values is 0.82. In the case of one mutation with a higher absolute AVG value than 0.82, it means that this mutation has a significant effect on protein activity or thermostability. In summary, most of the residues (30 out of 38) have high absolute AVG values either in activity or thermostability, indicating that SCA.SIM method has powerful predictive ability. For example, the predicted sites 129, 188, 223, and 225 are involved in catalytic activity and also observed in previous works.[31,53]

To further compare the performance between SCA and SCA.SIM, another 14 positions predicted by SCA and not covered by SCA.SIM were mutated to cysteine as shown in Figure S4A-4B. The results show these 14 mutants have little effects on bioactivity. This demonstrates that the performance of SCA.SIM is better than that of SCA. As control, we randomly selected 33 non-covariant sites (about 10% residues of CalB) and mutated them to cysteine (Figure S4C-4D). Similar results were found with 14 uncovered mutants. This further supports that the predicted covariant sites with SCA.SIM have important functions.

3.5 | SCA.SIM as a tool for protein engineering

From the data matrix of activity and thermostability, we can also found some good mutants, which means this method can also be used as a tool for protein engineering. Therefore, we constructed and characterized some activity enrichment mutants: A225M, W104V, M129L, and D223R. Using *p*NP

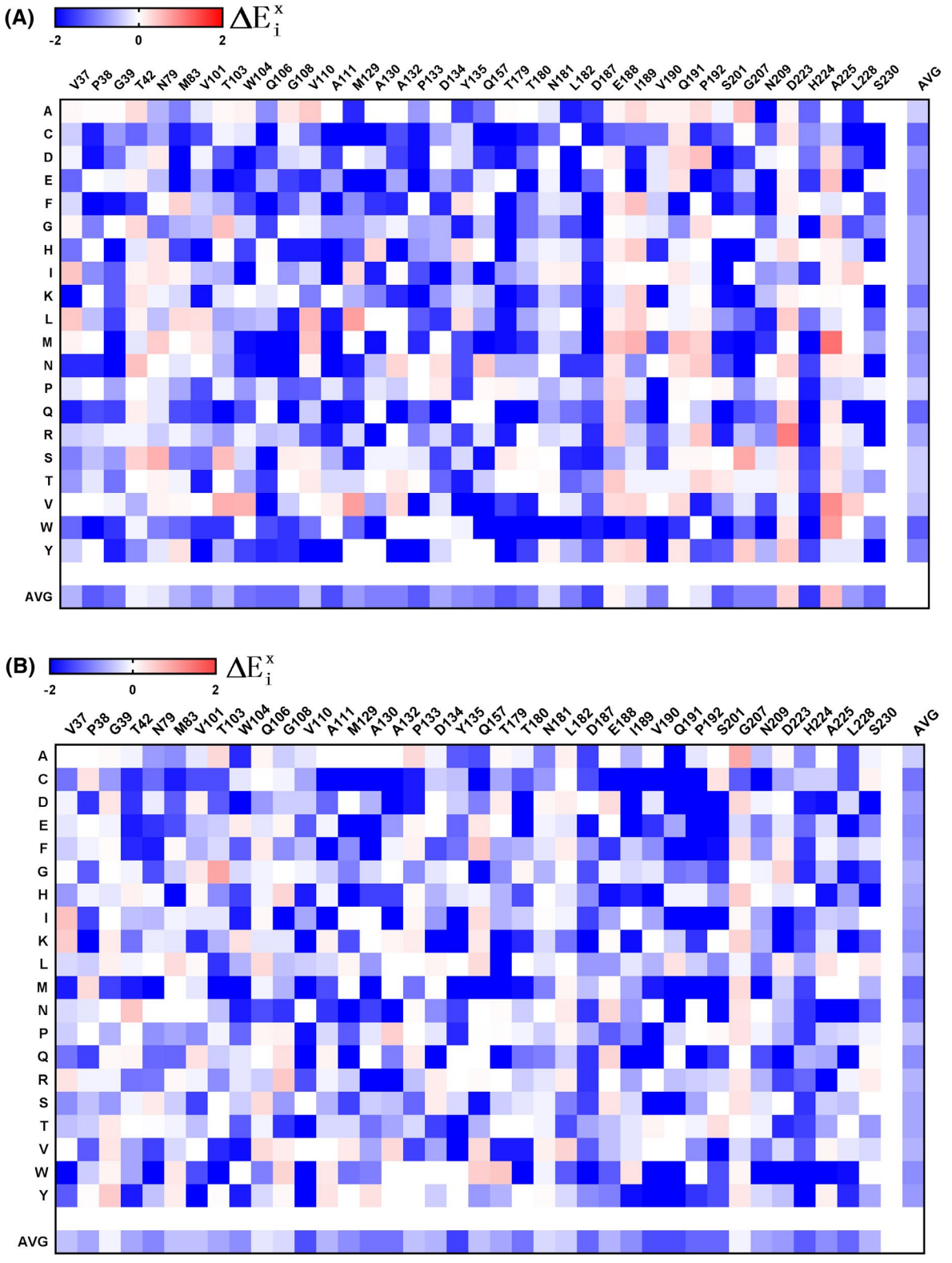

FIGURE 7 Complete single mutagenesis of covariant sites in CalB. A, Data matrix showing ΔE_i^x. –the activity effect of every mutation x at each position i relative to the wild type—colorimetrically, with blue representing loss-of-activity and red representing gain-of-activity mutations. B, Data matrix showing ΔE_i^x–the thermostability effect of every mutation x at each position i relative to the wild type—colorimetrically, with blue representing loss-of-thermostability and red representing gain-of-thermostability mutations

caprylate as the substrate, we measured the kinetic parameters of wild-type CalB and these mutants. All mutants demonstrated a major increase in k_{cat}, and the mutant A225M displayed the highest catalytic efficiency (k_{cat}/K_m), nearly 12.0-fold higher than that of the wild type and even higher than that of W104V, which showed a 9.4-fold increase (Table 1). To monitor the dynamic character of catalytic efficiency, 50 ns molecular dynamics simulations were performed for wild type and A225M. The binding free energy between A225M and substrate was lower than that of wild type and consistent with the previous measured kinetic parameters (Table S3). The mutation A225M induced significant conformation change at the opposite α-helix, which forms more hydrophobic interactions between enzyme and substrate as shown in Figures S5-S6. The distance between α-helix and the substrate is about 4 Å smaller in A225M than that in wild type. This can partly explain A225M with high catalytic efficiency. Therefore, we can suppose that these sites might play some key roles in protein activity and thermostability based on the complexity of interactions between amino acid residues.

As for the thermostability, the mutation sites 130, 180, and 228 show effects on the thermostability of CalB that are consistent with our predictions. In particular, our previous work constructed mutant D223G and found that it increased the thermostability and maintained the activity of the protein.[53] We can see the same results in the matrix (Figure 7B). To further evaluate the thermostability of positive mutants, V37I, T42N, T103G, and G207A were constructed. The half-life $t_{1/2}$, transition state free energy ($\Delta\Delta G$), and inactivation constant k_d of CalB are listed in Table 2. The $t_{1/2}$ of T103G at 50°C was 51 min, which was approximately sixfold higher than that of the wild type. Thermodynamic analysis showed that the $\Delta\Delta G$ values of V37I, T42N, T103G, and G207A were increased by 0.47, 0.2, 4.62, and 2.24 KJ·mol^{-1}, respectively. Their catalytic efficiencies (k_{cat}/K_m) were also increased. In particular, that of the mutant T103G increased approximately 11-fold. To evaluate the allosteric coupling which provides information on the energetic coupling between pairs of residues, we created a saturated mutant library (~3000 colonies) for a pair sites of A225 and T103. The best mutant A225F/T103A shows the activity about 9-fold and 20-fold than those of two corresponding single mutants (Table 1). This can give a hint that pairwise mutant might create higher activity than single mutant; therefore, we will continue to screen double mutants in the future work according to the specific pairwise correlation information.[54]

TABLE 1 Kinetic constants of wild-type and mutant CalB

Enzyme	K_m[a] (μM)	k_{cat} (min^{-1})	k_{cat}/K_m (min^{-1}μM^{-1})
WT	15 ± 0.3	675 ± 10	45 ± 1.2
A225M	11 ± 0.6	5623 ± 173	510 ± 13
W104V	22 ± 0.7	9355 ± 145	425 ± 7.5
D223R	7.3 ± 0.3	2339 ± 82	322 ± 23
M129L	11.2 ± 0.4	1313 ± 36	117 ± 5.7
T103G	7 ± 0.5	3522 ± 150	503 ± ± 14
G207A	13 ± 0.6	5716 ± 212	442 ± 35
T42N	6.7 ± 0.2	751 ± 18	113 ± 3.7
V37I	7.7 ± 0.1	1079 ± 45	140 ± 6
A225V	11.4 ± 0.3	2416 ± 66	212 ± 4.7
M129V	16 ± 0.5	1094 ± 31	68 ± 3.7
A225F	14 ± 0.6	2151 ± 50	152 ± 10
T103A	13 ± 0.3	921 ± 18	71 ± 1.7
A225F/ T103A	9.5 ± 0.4	13517 ± 230	1431 ± 87

Values represent the mean of three independent sets of experiments.

[a]The kinetic constants were determined at 37°C using *p*-NP caprylate as the substrate.

4 | DISCUSSION

4.1 | Comparison with SCA.MD

To evaluate the difference between SCA.MD[13] and SCA.SIM, the 2182 subfamily sequences used in SCA.SIM were realigned with Clustal Omega [55,56] based on sequence. The alignments based on sequence of SCA.MD and structure of SCA.SIM are significantly different (Figure S7). Then, SCA.MD matrix and the corresponding covariant residues were calculated using the same method as SCA.SIM. The probability distribution of the eigenmode values of the first eigenvector of SCA.MD is decentralized which may cause false selection (Figure 8A). As SCA.MD is only based on the sequence alignment, the calculated residues are more sequence connected than spatial structure connected (Figure 8B). We identified 45 covariant residues by SCA.MD, and 12 of them are annotated sites, which means SCA.MD also has a powerful predicted ability on the functional residues. However, the distribution of these 12 annotated sites is relatively dispersed. We statistics the overlap functional residues

TABLE 2 Kinetic stability properties of wild-type and mutant CalB

Enzyme	k_d[a] (min^{-1})	$t_{1/2}$[b] (min)	$\Delta\Delta G$[c] (KJ mol^{-1})
WT	0.081 ± 0.002	8.5 ± 0.3	–
V37I	0.068 ± 0.001	10.2 ± 0.4	0.47 ± 0.02
T42N	0.069 ± 0.003	9.2 ± 0.1	0.20 ± 0.01
T103G	0.014 ± 0.001	51.1 ± 1.3	4.62 ± 0.03
G207A	0.034 ± 0.002	20.3 ± 0.2	2.24 ± 0.02

–, not detected.

[a]k_d denotes the first-order rate constants of inactivation at 50°C.

[b]$t_{1/2}$ represents the half-life at 50°C and is equal to ln 2/k_d.

[c]$\Delta\Delta G = -RT \ln(k_d \text{ mutant}/k_d \text{ wild type})$. Values represent the mean of three independent sets of experiments.

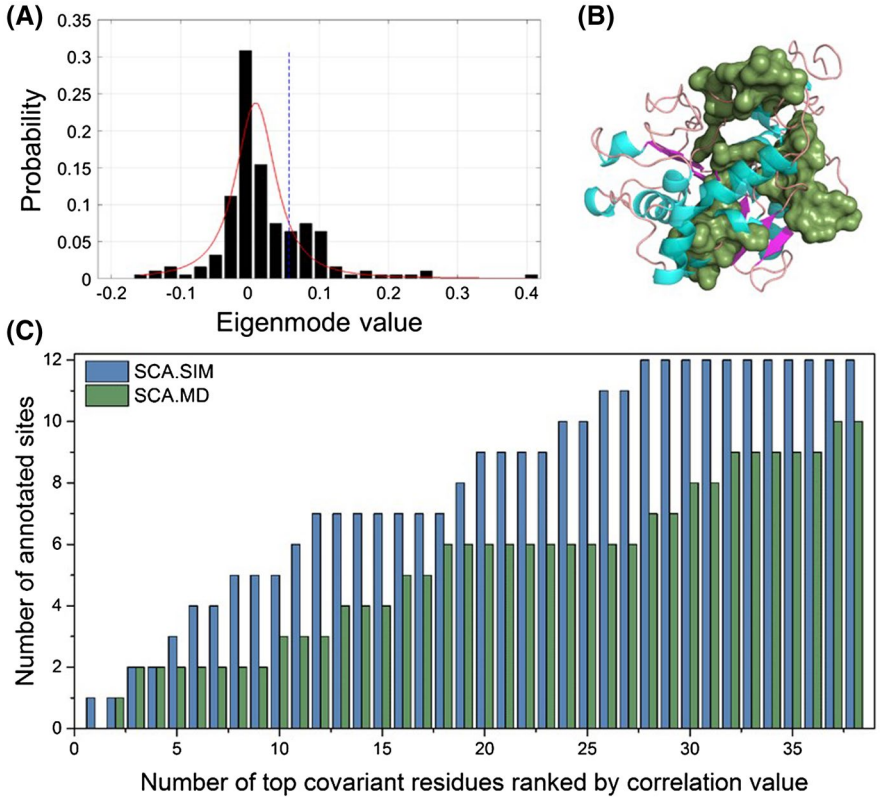

FIGURE 8 Results of SCA.MD method on CALB. A, The probability distribution analysis of the first eigenmode values of the SCA.MD. B, Covariant residues obtained for the tertiary structure of CALB by SCA.MD. C, The number of known functional residues among the covariant residues obtained by SCA.SIM and SCA.MD

predicted by SCA.MD and SCA.SIM (top 38) and the results are shown in Figure 8C. The number of covered known functional residues is 12 in SCA.SIM and 10 in SCA.MD. This indicates that SCA.SIM performs better than SCA.MD. And we mutated the other 31 SCA.MD sites without covered by SCA.SIM and the literature to cysteine as shown in Figure S8 which shows these mutants have little effects on bioactivity. This demonstrates that the performance of SCA.SIM is better than that of SCA.MD.

4.2 | SCA.SIM reveals multiple functions of covariant residues

Sectors have been proposed by Dr. Rama Ranganathan defined as groups of covariant residues in a protein family. They are characterized by statistical independence, physical connectivity within the tertiary structure, biochemical independence in mediating protein function, and independent phenotypic variation within the protein family. Using the S1A family as a model system, they identify three biochemical function independent sectors.[8] However, we found a new phenomenon for covariant residues with multiple functions in the α/β hydrolase subfamilies. That might be caused by the structure diversity among different enzyme families. And the complexity of these covariant residues also makes it difficult to interpret their biological function specifically. This implies that strict covariant residues independence need not be guaranteed in every protein family.[8] According to our results, the covariant residues are functionally important and are associated with activity and thermostability. In particular, the mutants G207A and T103G increase both activity and thermostability. Our results are in consistent with the previous works that mutate residues near active center improve activity and thermostability.[57] Therefore, we assumed that 38 covariant residues might modulate both functions, which gives a new sight for the diverse co-evolutionary routes.

4.3 | SCA.SIM provides a novel tool for enzyme evolution

Compared to traditional site-directed mutagenesis, SCA.SIM-guided protein engineering could identify more specific mutation sites for enzyme evolution. Due to the limitations of traditional protein engineering methods, most site-directed mutations are focused on the substrate binding pocket or the protein surface.[31,58] And these sites have been thoroughly studied and new functional sites are urgently needed. However, covariant residues are not confined to the binding pocket. The covariant residues from SCA.SIM can distribute inside the fundamental structure of protein. These residues are sites with the potential for strong impacts on enzyme function. Combination with the pairwise mutant strategy,[9] we can rapidly improve the function in covariant

sites. Therefore, SCA.SIM might provide a novel method for enzyme evolution. In summary, these results have enhanced the available methods for coevolution calculation and may provide a reliable reference for identifying the allosteric coupling and the directed evolution of the enzyme.

ACKNOWLEDGMENTS

This work was supported by Center for HPC at Shanghai Jiao Tong University, the National Key Research and Development Program of China (2017YFE0103300), the National Natural Science Foundation of China (31770771 and 31620103901), and the Medical Engineering Cross Fund of Shanghai Jiao Tong University (YG2017MS08).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Zhiyun Wu conducted the experiments, analyzed the results, and wrote most of the paper. Hao Liu and Lishi Xu completed algorithm optimization and contributed to the preparation of the figures. Haifeng Chen and Yan Feng conceived the idea for the project, designed the experiment, and analyzed the data. All authors analyzed the results and approved the final version of the manuscript.

REFERENCES

1. Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D. Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci*. 2017;114:9122-9127.
2. Bai F, Morcos F, Cheng RR, Jiang H, Onuchic JN. Elucidating the druggable interface of protein− protein interactions using fragment docking and coevolutionary analysis. *Proc Natl Acad Sci*. 2016;113:E8051-E8058.
3. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci*. 2011;108:E1293-E1301.
4. Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nature Reviews Genetics*. 2013;14(4):249–261.
5. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. *Nature*. 2005;437:512-518.
6. Lee J, Natarajan M, Nashine VC, et al. Surface sites for engineering allosteric control in proteins. *Science*. 2008;322:438-442.
7. Liu Y, Yan Z, Lu X, Xiao D, Jiang HJS. Improving the catalytic activity of isopentenyl phosphate kinase through protein coevolution analysis. *Sci Rep*. 2016;6:24117.
8. Halabi N, Rivoire O, Leibler S, Ranganathan RJC. Protein sectors: evolutionary units of three-dimensional structure. *Cell*. 2009;138:774-786.
9. Salinas VH, Ranganathan RJE. Coevolution-based inference of amino acid interactions underlying protein function. *eLife*. 2018;7:e34300.
10. Reynolds KA, McLaughlin RN, Ranganathan R. Hot spots for allosteric regulation on protein surfaces. *Cell*. 2011;147:1564-1575.
11. McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. *Nature*. 2012;491:138–144.
12. Jacob E, Unger R, Horovitz AJE. Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. *Elife*. 2015;4:e08932.
13. Estabrook RA, Luo J, Purdy MM, et al. Statistical coevolution analysis and molecular dynamics: identification of amino acid pairs essential for catalysis. *Proc Natl Acad Sci*. 2005;102:994-999.
14. Kuipers RK, Joosten H-J, van Berkel WJH, et al. 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Prot: Struc Funct Bioinformatics* 2010;78:2101-2113.
15. Kourist R, Jochens H, Bartsch S, et al. The α/β-Hydrolase fold 3DM database (ABHDB) as a tool for protein engineering. *ChemBioChem*. 2010;11:1635-1643.
16. Uppenberg J, Hansen MT, Patkar S, Jones TA. The sequence, crystal structure determination and refinement of two crystal forms of lipase B from Candida antarctica. *Structure*. 1994;2:293-308.
17. Tosco P, Balle T. Open3DQSAR: a new open-source software aimed at high-throughput chemometric analysis of molecular interaction fields. *J Mol Model*. 2011;17:201-208.
18. Case DA, Berryman JT, Betz RM, et al. *AMBER 2015*. San Francisco: University of California; 2015.
19. Darden T, York D, Pedersen L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J Chem Phys*. 1993;98:10089-10092.
20. Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*. 2006;25:247-260.
21. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*. 1977;23:327-341.
22. Tai K, Shen T, Börjesson U, Philippopoulos M, McCammon JA. Analysis of a 10-ns molecular dynamics simulation of mouse acetylcholinesterase. *Biophys J*. 2001;81:715-724.
23. Young MA, Gonfloni S, Superti-Furga G, Roux B, Kuriyan J. Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell*. 2001;105:115-126.
24. Hünenberger P, Mark A, Van Gunsteren W. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J Mol Biol*. 1995;252:492-503.
25. Ichiye T, Karplus M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Prot: Struct Funct, Bioinformatics*. 1991;11:205-217.
26. Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA: protein complexes. *Proc Natl Acad Sci*. 2009;106:6620-6625.
27. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504.
28. Floyd RW. Algorithm 97: shortest path. *Commun ACM*. 1962;5:345.
29. Stone J, Developers NT, Eargle J, Sethi A, Li L, Luthey-Schulten Z. Dynamical Network Analysis; 2012. Champaign: University of Illinios at Urbana.
30. Kim SJ, Lee JA, Joo JC, Yoo YJ, Kim YH, Song BK. The development of a thermostable CiP (Coprinus cinereus peroxidase) through in silico design. *Biotechnol Prog*. 2010;26:1038-1046.

31. Wu Q, Soni P, Reetz MTJJ. Laboratory evolution of enantiocomplementary *Candida antarctica* lipase B mutants with broad substrate scope. *J. Am. Chem. Soc*. 2013;135:1872-1881.
32. Rotticci D, Rotticci-Mulder JC, Denman S, Norin T, Hult KJC. Improved enantioselectivity of a lipase by rational protein engineering. *ChemBioChem*. 2001;2:766-770.
33. Engstrom K, Vallin M, Syren P-O, Hult K, Backvall J-E. Mutated variant of *Candida antarcticalipase* B in (S)-selective dynamic kinetic resolution of secondary alcohols. *Org Biomol Chem*. 2011;9:81-82.
34. Fuentes G, Ballesteros A, Verma CS. Specificity in lipases: a computational study of transesterification of sucrose. *Protein Sci*. 2004;13:3092-3103.
35. Hamberg A, Maurer S, Hult K. Rational engineering of *Candida antarctica* lipase B for selective monoacylation of diols. *Chem Commun*. 2012;48:10013-10015.
36. Hong S, Yoo Y. Activity enhancement of *Candida antarctica* lipase B by flexibility modulation in helix region surrounding the active site. *Appl Biochem Biotechnol*. 2013;170:925-933.
37. Irani M, Törnvall U, Genheden S, Larsen MW, Hatti-Kaul R, Ryde U. Amino acid oxidation of *Candida antarctica* lipase B studied by molecular dynamics simulations and site-directed mutagenesis. *Biochemistry*. 2013;52:1280-1289.
38. Juhl PB, Doderer K, Hollmann F, Thum O, Pleiss J. Engineering of *Candida antarctica* lipase B for hydrolysis of bulky carboxylic acid esters. *J Biotechnol*. 2010;150:474-480.
39. Jung S, Park S. Improving the expression yield of *Candida antarctica* lipase B in *Escherichia coli* by mutagenesis. *Biotechnol Lett*. 2008;30:717-722.
40. Le QAT, Joo JC, Yoo YJ, Kim YH. Development of thermostable *Candida antarctica* lipase B through novel in silico design of disulfide bridge. *Biotechnol Bioeng*. 2012;109:867-876.
41. Linder M, Hermansson A, Liebeschuetz J, Brinck T. Computational design of a lipase for catalysis of the Diels-Alder reaction. *J Mol Model*. 2011;17:833-849.
42. Liu D, Trodler P, Eiben S, et al. Rational design of pseudozyma antarctica lipase B yielding a general esterification catalyst. *ChemBioChem*. 2010;11:789-795.
43. Park C-G, Kwon M-A, Song J-K, Kim D-M. Cell-free synthesis and multifold screening of *Candida antarctica* lipase B (CalB) variants after combinatorial mutagenesis of hot spots. *Biotechnol Prog*. 2011;27:47-53.
44. Park HJ, Joo JC, Park K, Kim YH, Yoo YJ. Prediction of the solvent affecting site and the computational design of stable *Candida antarctica* lipase B in a hydrophilic organic solvent. *J Biotechnol*. 2013;163:346-352.
45. Peng X-Q. Improved thermostability of lipase B from *Candida antarctica* by directed evolution and display on yeast surface. *Appl Biochem Biotechnol*. 2013;169:351-358.
46. Skjøt M, De Maria L, Chatterjee R, et al. Understanding the plasticity of the α/β hydrolase fold: lid swapping on the *Candida antarctica* lipase B results in chimeras with interesting biocatalytic properties. *ChemBioChem*. 2009;10:520-527.
47. Tanino T, Ohno T, Aoki T, Fukuda H, Kondo A. Development of yeast cells displaying *Candida antarctica* lipase B and their application to ester synthesis reaction. *Appl Microbiol Biotechnol*. 2007;75:1319-1325.
48. Vallin M, Syrén P-O, Hult K. Mutant lipase-catalyzed kinetic resolution of bulky phenyl alkyl sec-alcohols: a thermodynamic analysis of enantioselectivity. *ChemBioChem*. 2010;11:411-416.
49. Zhang N, Suen WC, Windsor W, Xiao L, Madison V, Zaks A. Improving tolerance of *Candida antarctica* lipase B towards irreversible thermal inactivation through directed evolution. *Protein Eng*. 2003;16:599-605.
50. Adkar BV, Tripathi A, Sahoo A, et al. Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure*. 2012;20:371-381.
51. Fowler DM, Araya CL, Fleishman SJ, et al. High-resolution mapping of protein sequence-function relationships. *Nat Methods*. 2010;7:741-746.
52. Van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods*. 2009;6:767-772.
53. Xie Y, An J, Yang G, et al. Enhanced enzyme kinetic stability by increasing rigidity within the active site. *J. Biol. Chem.* 2014;289:7994-8006.
54. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife*. 2018;7:e34300.
55. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 2011;7: 539–544.
56. Sievers F, Higgins DGJPS. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 2018;27:135-145.
57. Sun Z, Liu Q, Qu G, Feng Y, Reetz MT. Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem Rev*. 2019;119:1626-1665.
58. Yi Z-L, Pei X-Q, Wu Z-L. Introduction of glycine and proline residues onto protein surface increases the thermostability of endoglucanase CelA from Clostridium thermocellum. *Bioresour Technol*. 2011;102:3636-3638.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wu Z, Liu H, Xu L, Chen H-F, Feng Y. Algorithm-based coevolution network identification reveals key functional residues of the α/β hydrolase subfamilies. *The FASEB Journal*. 2020;34:1983–1995. https://doi.org/10.1096/fj.201900948RR