ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

# Mutation Space of Spatially Conserved Amino Acid Sites in Proteins

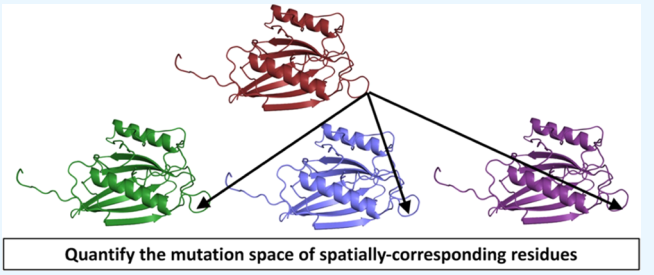
Benjamin T. Caswell,[†] Thomas J. Summers,[†] Gerra L. Licup, and David C. Cantu*

Cite This: *ACS Omega* 2023, 8, 24302−24310

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** The mutation space of spatially conserved (MSSC) amino acid residues is a protein structural quantity developed and described in this work. The MSSC quantifies how many mutations and which different mutations, i.e., the mutation space, occur in each amino acid site in a protein. The MSSC calculates the mutation space of amino acids in a target protein from the spatially conserved residues in a group of multiple protein structures. Spatially conserved amino acid residues are identified based on their relative positions in the protein structure. The MSSC examines each residue in a target protein, compares it to the residues present in the same relative position in other protein structures, and uses physicochemical criteria of mutations found in each conserved spatial site to quantify the mutation space of each amino acid in the target protein. The MSSC is analogous to scoring each site in a multiple sequence alignment but in three-dimensional space considering the spatial location of residues instead of solely the order in which they appear in a protein sequence. MSSC analysis was performed on example cases, and it reproduces the well-known observation that, regardless of secondary structure, solvent-exposed residues are more likely to be mutated than internal ones. The MSSC code is available on GitHub: "https://github.com/Cantu-Research-Group/Mutation_Space".




## 1. INTRODUCTION

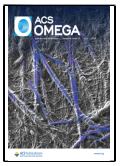
A fundamental question about protein structure is how the three-dimensional structure of a protein will change by specific mutations. This work presents a way to quantify how many mutations are present in the spatial location of an amino acid site in the tertiary structure of a protein as a metric to determine which mutations are more likely to affect the protein structure.

Proteins can be compared based on their sequence to find similar proteins among all known sequences,[1,2] which has enabled countless advances in protein science and engineering because proteins with similar sequences are likely to have similar structures and functions. An approach to predict how mutations affect protein function is through examination of sequences and comparison to function, for example, focusing on single-nucleotide polymorphism. Amino acid sequence-based approaches tend to not rely on direct protein tertiary structure comparisons, instead making function predictions by comparing available sequencing data.[3−8] Although sequence data alone may be sufficient to predict function, generally sequences alone are not sufficient to make specific structural predictions: for example, in homology modeling, sequence comparisons identify known structures that serve as templates to predict three-dimensional structures.[9,10]

Proteins can be compared by superimposing their three-dimensional structures. Tertiary structure superimposition, or alignment, methods provide consistent and useful information for comparing the protein structure through direct tertiary structural comparison[11,12] and allow making inferences on how specific amino acids affect function. To compare highly divergent proteins that may share only a small conserved core or region, structural comparisons can be performed by superimposing only highly similar structural fragments.[13] Superimposition approaches focus on attaining the best global fit by minimizing the distance between residues or sub-structures in different proteins. Tertiary structure comparison provides data that sequence-driven methods struggle to attain, but, like sequence-based comparison methods that seek to optimize the local or global alignment, tertiary structure alignment methods minimize distances between structures to optimally superimpose protein structures.

In this work, we present the mutation space of spatially conserved (MSSC) amino acid sites in proteins. The MSSC compares protein tertiary structures by identifying spatially corresponding residues based on their relative positions in their protein structures. This method quantifies how many mutations occur in each spatially conserved amino site in a target protein in a group of multiple protein structures. The MSSC examines each residue in a target protein, compares it to the residues present in the same relative position in other protein structures from that group, and uses the physicochem-

Received: March 4, 2023
Accepted: June 14, 2023
Published: June 28, 2023



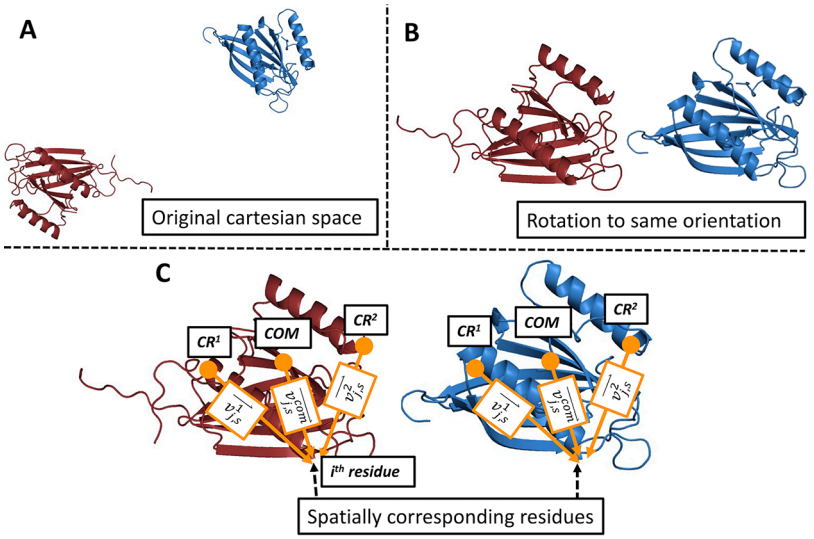



**Figure 1.** Visualization of spatially corresponding residues in two structures. (A) Protein structures are in the cartesian space, for example, in a Protein Data Bank structure. (B) Structures are oriented in cartesian space according to their reference points independently with no information from other structures. (C) Spatially corresponding residues, which occupy the same relative spatial positions in their structures, can be determined by matching the magnitude and orientation of the three vectors in eqs 1−3.

ical criteria of mutations found in each conserved spatial site to quantify the mutation space of that residue. The MSSC provides a unique perspective because it does not seek to identify the best sequence fit or the best structural superimposition but rather only informs how many amino acid mutations occur in a spatial site in a protein structure. The MSSC of each residue in a protein is simply a score of how many different amino acid residues appear in a spatially conserved site.

## 2. RESULTS

The mutation space of a spatially conserved amino acid site is a quantification of the overall conservation of a specific spatial location for a group of similar proteins, accounting for the diversity of amino acids found in a particular site and the degree of spatial and physicochemical conservation of that site. This is analogous to scoring conserved sites in a multiple sequence alignment but in three-dimensional space considering the spatial location of residues, instead of solely the order in which they appear in a protein sequence.

To be able to quantify the mutation space for a site in three-dimensional space for several proteins, corresponding residues for a spatial site must be identified. At most, one residue from each protein in a group of protein structures can occupy a site in three-dimensional space: these are the spatially corresponding residues, and how they are identified is described in Section 2.1. Once spatially corresponding residues are identified, the mutation space is calculated based on the residues that appear in each spatial site in three-dimensional space, as described in Section 2.2.

**2.1. Identifying Spatially Corresponding Residues in the Tertiary Structure of Proteins.** A method to identify spatially corresponding residues was developed in this work to avoid relying on protein structure superimposition approaches and external software, since the goal is not to superimpose structures with the best overlap, but rather to identify spatially corresponding residues between two protein structures based on the relative position of each residue in its protein structure. Within a set of structures (i.e., a group of proteins), a target protein structure is selected, while the remaining ones are the subject protein structures. All of the amino acid sequences in the group of proteins are used to obtain a multiple sequence alignment. Sequence-conserved residues (i.e., identities, commonly denoted with a * in a multiple sequence alignment) as well as nearly-conserved residues (i.e., similarities, commonly denoted with a : in a multiple sequence alignment) are identified from the multiple sequence alignment. For each three-dimensional structure in the group of proteins, the average position of sequence-conserved residues is labeled as the center of mass of conserved residues for that protein structure (**COM**), which is then defined as its origin in cartesian space. For each three-dimensional structure in the group of proteins, the two sequence-conserved residues (i.e., identities) in the set that are separated by the greatest spatial distance are then selected as reference points, $\mathbf{CR^1}$ and $\mathbf{CR^2}$, keeping assignments consistent between each protein structure in the group. When there are no identities with spatial distance in the multiple sequence alignment, nearly-conserved residues (i.e., similarities) are then used to select $\mathbf{CR^1}$ and $\mathbf{CR^2}$. Each protein structure is then rotated about its **COM** origin such that $\mathbf{CR^1}$ is aligned with the $z$-axis and $\mathbf{CR^2}$ is on the $\{(x, 0, z)|x \geq 0, z \in \mathbb{R}\}$ plane.

Following this spatial realignment, the position of each residue $j$ within a protein structure $s$, $r_{j,s}$, is redefined by the vectors

$$\overrightarrow{v^1_{j,s}} = \overrightarrow{\mathrm{CR}^1_s r_{j,s}} \quad (1)$$

$$\overrightarrow{v^2_{j,s}} = \overrightarrow{\mathrm{CR}^2_s r_{j,s}} \quad (2)$$

$$\overrightarrow{v^{\mathrm{com}}_{j,s}} = \overrightarrow{\mathrm{COM}_s r_{j,s}} \quad (3)$$

The initial spatial realignment and vector conversion results in the protein structures that are highly similar will have highly similar vector fields, regardless of the absolute position and orientation of each protein structure in cartesian space. Each



https://doi.org/10.1021/acsomega.3c01473
ACS Omega 2023, 8, 24302−24310

residue from a subject protein structure ***s***, $r_{j,s}$, is compared to each residue from the target protein structure ***t***, $r_{i,t}$.

The similarity of the spatial positions of two residues is determined through examination of related defining vectors, see Figure 1. The position of each residue is defined by three vectors, which originate at the selected reference points, **CR**[1], **CR**[2], and **COM**. These reference points, for the groups of similar proteins studied, have highly conserved spatial positions in all structures. Therefore, if the vectors defining a residue in a subject structure, $\overrightarrow{v^1_{j,s}}$, $\overrightarrow{v^2_{j,s}}$, and $\overrightarrow{v^{com}_{j,s}}$, are all oriented in the same direction and have the same magnitudes as the vectors defining a residue in the target structure, $\overrightarrow{v^1_{i,t}}$, $\overrightarrow{v^2_{i,t}}$, and $\overrightarrow{v^{com}_{i,t}}$, respectively, then the two residues must occupy the same spatial position relative to their reference points.

To quantify the similarity of vector orientation, the triangle similarity-section similarity (TS-SS)[14] approach was used to calculate the similarity between the subject structure vectors ($\vec{v}^1_{j,s}$, $\vec{v}^2_{j,s}$, $\vec{v}^{com}_{j,s}$) and the corresponding vectors in the target structure ($\vec{v}^1_{i,t}$, $\vec{v}^2_{i,t}$, $\vec{v}^{com}_{i,t}$) for all ***i***th and ***j***th residues. TS-SS is a geometric similarity measure that combines elements from cosine, Euclidian distance, and magnitude difference metrics to distinguish vectors. The TS between the target and subject vectors is calculated from the vector magnitudes and the angle between the vectors:

$$TS^1 = TS(\vec{v}^1_{j,s}, \vec{v}^1_{i,t}) = \frac{\| \vec{v}^1_{j,s} \| \cdot \| \vec{v}^1_{i,t} \| \cdot \sin(\theta)}{2} \quad (4)$$

$$TS^2 = TS(\vec{v}^2_{j,s}, \vec{v}^2_{i,t}) = \frac{\| \vec{v}^2_{j,s} \| \cdot \| \vec{v}^2_{i,t} \| \cdot \sin(\theta)}{2} \quad (5)$$

$$TS^{com} = TS(\vec{v}^{com}_{j,s}, \vec{v}^{com}_{i,t}) = \frac{\| \vec{v}^{com}_{j,s} \| \cdot \| \vec{v}^{com}_{i,t} \| \cdot \sin(\theta)}{2} \quad (6)$$

The SS between the target and subject vectors is similarly computed as

$$SS^1 = SS(\vec{v}^1_{j,s}, \vec{v}^1_{i,t}) = \pi \cdot \left(\sqrt{(\vec{v}^1_{j,s} - \vec{v}^1_{i,t})^2} + |\| \vec{v}^1_{j,s} \| - \| \vec{v}^1_{i,t} \||\right)^2 \cdot \left(\frac{\theta}{360}\right) \quad (7)$$

$$SS^2 = SS(\vec{v}^2_{j,s}, \vec{v}^2_{i,t}) = \pi \cdot \left(\sqrt{(\vec{v}^2_{j,s} - \vec{v}^2_{i,t})^2} + |\| \vec{v}^2_{j,s} \| - \| \vec{v}^2_{i,t} \||\right)^2 \cdot \left(\frac{\theta}{360}\right) \quad (8)$$

$$SS^{com} = SS(\vec{v}^{com}_{j,s}, \vec{v}^{com}_{i,t}) = \pi \cdot \left(\sqrt{(\vec{v}^{com}_{j,s} - \vec{v}^{com}_{i,t})^2} + |\| \vec{v}^{com}_{j,s} \| - \| \vec{v}^{com}_{i,t} \||\right)^2 \cdot \left(\frac{\theta}{360}\right) \quad (9)$$

where $\theta$ is the angle between the two vectors that appear in each equation in eqs 4−9. TS-SS multiplies the area of the triangle between the two vectors and the area of the sector with a radius of the sum of the Euclidian distance and magnitude difference of the two vectors. The TS and SS quantities are combined into a TS-SS value

$$TS - SS = \frac{TS^1 \cdot SS^1 + TS^2 \cdot SS^2 + TS^{com} \cdot SS^{com}}{3} \quad (10)$$

Using this approach, the TS-SS value will range from 0 to ∞, where identical vectors will have a TS-SS value of 0 and it will increase as the difference between vector direction and magnitude increases. For each subject−target residue pair, the TS-SS value between the defining target and subject vectors is used to measure residue spatial similarity. Testing of the method on the structure datasets discussed in this work identified that a TS-SS value threshold of 1.0 has sufficient flexibility to capture spatially correlated residues.

The subject−target analysis to identify which residue in the subject structure spatially corresponds with a residue in the target structure is run for every pair of subject−target protein structures, where the subject proteins are proteins in the group that are not the target. For each residue in the target protein, a list of possible subject residue matches is generated for each subject protein. For a given subject−target protein pair, each target residue may be possibly correlated to zero or more subject residues.

For a given subject−target pair, only one subject residue can be considered to occupy the same spatial position as a given target residue. For all residues in a target structure, the following algorithm was developed to make a final assignment of at most one subject residue from each subject structure to each target residue. First, subject−target residue pairs that are incredibly spatially similar (defined as having a TS-SS value of 0.03) to only each other are assigned as being spatially correlated, and those subject residues are removed as possible matches from the possible correlation set. After this initial assignment, target residues that have only one possible spatially correlated subject residue, according to the criteria described, are identified, and those subject residues are assigned as being spatially correlated to their respective target residues. Then, the subject residues in these 1:1 correlations are likewise removed as possible matches from the possible correlation set. This process is repeated until no more 1:1 (subject residue/target residue) matches are found. The second step identifies 2:2 correlations, meaning when two adjacent target residues are possibly spatially correlated to two adjacent subject residues, e.g., $r_{15,t}$ and $r_{16,t}$ are both possibly correlated to $r_{23,s}$ and $r_{24,s}$. These are assigned to minimize the total TS-SS value, e.g., $r_{23,s}$ is assigned to $r_{15,t}$ and $r_{24,s}$ is assigned to $r_{16,t}$ if

$$TS - SS_{15,23} + TS - SS_{16,24} < TS - SS_{16,23} + TS - SS_{15,24} \quad \text{(Exp 1)}$$

These subject residues are assigned as spatially correlated to target residues and removed from the possible correlation set as in the 1:1 cycle. This cycle is repeated until no more 2:2 matches are found. The 1:1 and 2:2 cycles are alternately repeated until no more changes are made to the assignment of spatially correlated residues. This macrocycle assigns most of the final target-subject residue pairs; the remaining residues are assigned in a manner, which retains the sequential alignment between the target and subject protein structures.

**2.2. Quantifying the Mutation Space in Spatially Conserved Sites.** For each residue in the target protein


https://doi.org/10.1021/acsomega.3c01473
ACS Omega 2023, 8, 24302−24310

structure, the spatially corresponding residues of each subject protein structure are identified (Section 2.1); now, the mutation space of each residue in the target protein structure needs to be quantified. In a set of structures (i.e., a group of proteins), there are $x+1$ structures in a set: one target structure against which $x$ subject structures are compared. A residue in any subject protein structure is a spatially conserved residue if it occupies the same location in three-dimensional space as a corresponding target protein residue, for both, relative to their respective structures. Each subject protein residue can be spatially correlated, or conserved, to at most one residue from the target protein structure. Also, it is possible that a target protein residue does not spatially correlate with any residues from a subject structure: if the criteria described in Section 2.1 is not met, then a target protein residue does not have a spatially corresponding subject protein residue.

To quantify the mutation space of each residue $i$ in the target protein structure, we developed the mutation space of spatially correlated residues ($\mathrm{MSSC}_i$)

$$\mathrm{MSSC}_i = \frac{n_{\mathrm{unique}}}{c} \sum_{1}^{x} \left( \sum_{r=1}^{19} D_{i,r} \right) \tag{11}$$

which is used to score each residue in the target structure. Each target protein residue can have at most $x$ correlated subject protein residues, where $x$ is the number of subject protein structures, see Figure 2. When a target residue is found to be spatially correlated to $c$ subject residues, at most one per subject protein structure, then there are $x - c$ instances of subject structures having no spatially correlated residue for that specific residue of the target protein. At the core of this equation is the Grantham's distance,[15] $D_{i,r}$, indicating the distance between target residue $i$ and one of the $r = 19$ other standard amino acids. For scoring purposes, any nonstandard amino acids are treated as their nearest relative, e.g., selenomethionine is scored as if it were methionine.

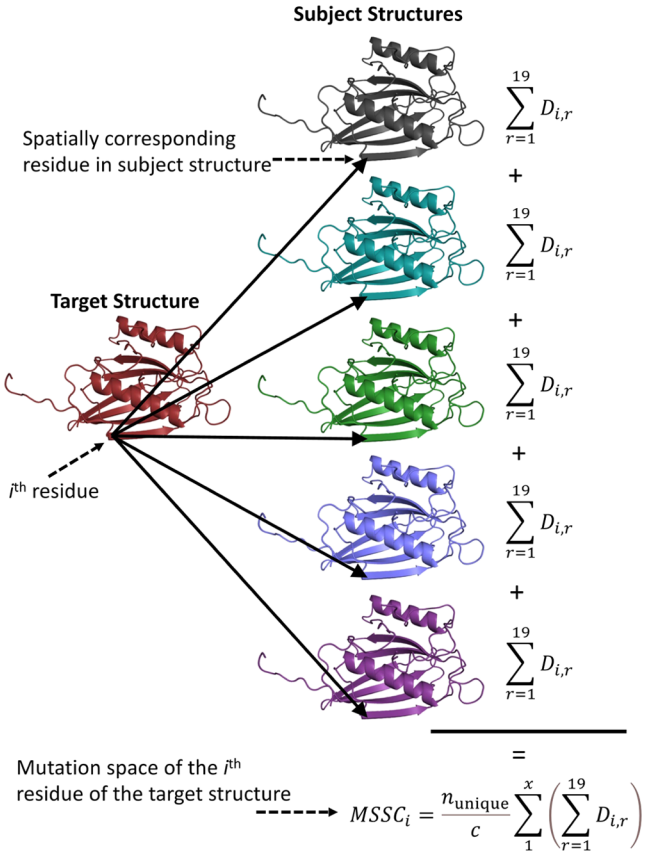



**Figure 2.** Visual representation of an MSSC calculation of a single residue in the target structure. The MSSC is calculated, using eqs 11 and 12, for every residue in the target structure based on the residues present in its spatially corresponding site in each subject structure.

To clarify, the MSSC is not a substitution matrix, rather a substitution matrix is used to score individual mutations, $D_{i,r}$. Grantham's distance is used because it is based purely on physicochemical criteria and does not use a sample of protein structures to optimize specific mutation scores. Different substitution matrices have been developed and optimized,[16−24] mostly with the goal of optimizing sequence alignments,[25] or to improve protein homology identification and the agreement between sequence and structure alignments.[26]

In the Grantham's distance matrix, there are $n_{i,r}$ instances of a given mutation, such as Cys−Trp or Cys−Glu. The total number of mutations on each target residue, $c$, is described by

$$c = \sum_{1}^{x} \left( \sum_{r=1}^{19} n_{i,r} \right) \tag{12}$$

and the number of unique mutations, $n_{\mathrm{unique}}$, gives the count of unique, nonequivalent amino acid correlations. For example, given a set of 11 proteins with one target structure and $x = 10$ subject structures, for a Cys residue in the target protein structure that is spatially correlated to 3 Cys in three different subject structures, 1 Ser in another subject structure, and 6 Gly in six different subject structures, it has $c = 10$ spatially correlated residues and $n_{\mathrm{unique}} = 2$ unique mutations (Cys−Gly, Cys−Ser). The mutational space score of residue $i$ in the target protein structure, $\mathrm{MSSC}_i$, is low for a target residue that has fewer and more similar mutations among its spatially correlated subject residues and high for a target residue that has many different mutations and more dissimilar mutations. The $\mathrm{MSSC}_i$ can also be low if $c$ is small $\left(\text{approx. } \frac{c}{x} = 0.5\right)$ since the scores for the $x - c$ noncorrelated subject residues are counted as zero. A convenient value to observe is the sample occurrence ratio

$$\mathrm{SOR} = \frac{c}{x} \tag{13}$$

which indicates the proportion of the subject protein structures that were found to have a correlated residue for a given target residue. Although the MSSC is focused on amino acid substitutions, insertion/deletions (indels) also have an effect on the structure of proteins in families.[27] The sample occurrence ratio (SOR) is a better indicator of indels, as a high number of indels would result in a lower number of spatially corresponding residues.

The mutation space code is available on GitHub (see the Methods section) to identify spatially corresponding residues and their mutation space. The code allows flexibility to users in two significant ways. First, users can turn off the structural alignment part of our code, if they prefer to calculate the mutation space with structures superimposed with a different method, for example, with Multiprot or TM-Align.[11,28] Second, users can calculate the mutation space with a different substitution matrix than Grantham's distance.


https://doi.org/10.1021/acsomega.3c01473
ACS Omega 2023, 8, 24302−24310

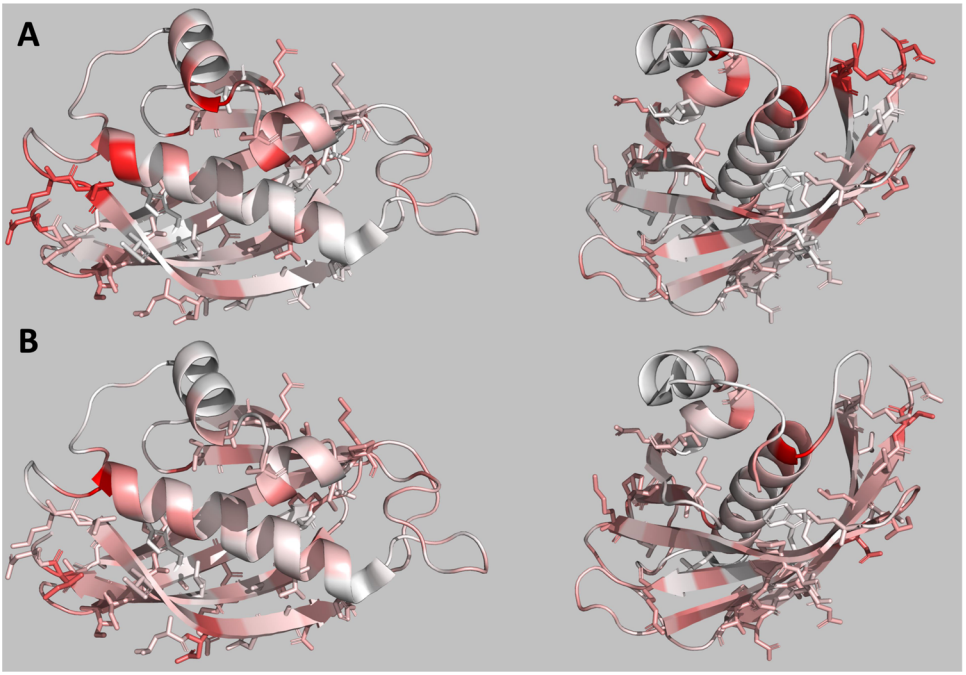



**Figure 3.** TE11 structure 1SC0 with a HotDog fold. Residues are colored red according to the MSSC score obtained when 1SC0 is grouped with (a) only proteins of the TE11 family or (b) proteins of both TE11 and TE6 families. Darker red indicates residues with a higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score for each grouping. Residue side chains as sticks are shown to highlight differences between interior and exterior residues.

The mutation space code is best suited to calculate the mutation space of residues in a protein within a group of proteins that have a substantial degree of structural and sequence similarity since the structural alignment approach has a limitation: at least two residues must be conserved, or nearly conserved, in the multiple sequence alignment ($\mathbf{CR}^1$ and $\mathbf{CR}^2$) in order to identify pairs of spatially corresponding residues. Therefore, to calculate the mutation space of the residues in a protein in a dissimilar group of proteins, the structural alignment part of our code would need to be turned off, as identifying spatially corresponding pairs needs some sequence and structural similarity.

## 3. DISCUSSION

To illustrate how MSSC analysis can lead to structural insights into proteins, we present four example cases from different protein folds and enzyme functions: two thioesterase enzymes, one ketoacyl synthase enzyme, and one glycoside hydrolase enzyme. For each case, the target and subject structures are listed in the Supporting Information. Although the cases presented are with experimentally resolved structures only, MSSC analysis could also be done with computationally predicted structures.

Examination of the mutation space of the four test cases reveals a common trend: mutations are more common and pronounced on the protein surface than the interior. This is observed for all cases with different structural folds, and enzymatic functions, reproducing the well-known observation that interior residues not in contact with the solvent play a role in maintaining the three-dimensional structure of proteins.[29−32] Exterior $\alpha$ helices display this phenomenon clearly; positions on an $\alpha$ helix that lie closer to the main bulk of the protein are consistently more conserved than positions that are more exposed to the solvent. This pattern is less pronounced in $\alpha$ helices that are "buried" in the structure, suggesting that the likelihood of solvent interaction may play a role in selecting/promoting mutations. In $\beta$ sheets, amino acids that are exposed to the solvent have a higher mutation space than those exposed to the protein structure core. Even in loops, their inherent disorder results in this pattern being less clear, but it is still present. Therefore, regardless of the secondary structure, solvent-exposed residues have a higher mutation space than internal ones; MSSC provides a metric to quantify this, although it has been known in protein science for decades that solvent-exposed residues are more likely to be mutated.[29−32]

Enzymes in the thioesterase (TE) family TE11[33] have a HotDog fold, and they hydrolyze acyl-CoA thioester bonds in many pathways, for example, enterobactin biosynthesis. Protein Data Bank[34] (PDB) structure 1SC0 is a TE11 *Haemophilus influenzae* enzyme and is the target structure for which $MSSC_i$ is calculated with respect to the other 37 TE11 structures, see Figure 3a. The MSSC of each residue in the target structure was calculated. The scores, both absolute and relative, can vary depending on the size and composition of the set of proteins studied, so it is convenient to visualize the results in relative terms. Therefore, in Figure 3a, the mutation space scores of residues in 1SC0 are expressed as a heatmap, where the darkest red represents the highest MSSC for that target structure. Inspection of the mutation space of 1SC0 shows a notable asymmetry: there is an orientation to the MSSC scores in secondary structures. Residues that lie on the exterior of the protein, or those that are more solvent-exposed, tend to have higher MSSC scores than those facing the interior. This phenomenon is clear in the $\beta$ sheet present in 1SC0: the 26 residues residing on the exterior surface have an average MSSC of 210 and SOR of 0.978, while the 23 interior residues have an average MSSC of 110 and SOR of 0.998.


https://doi.org/10.1021/acsomega.3c01473
*ACS Omega* 2023, 8, 24302−24310

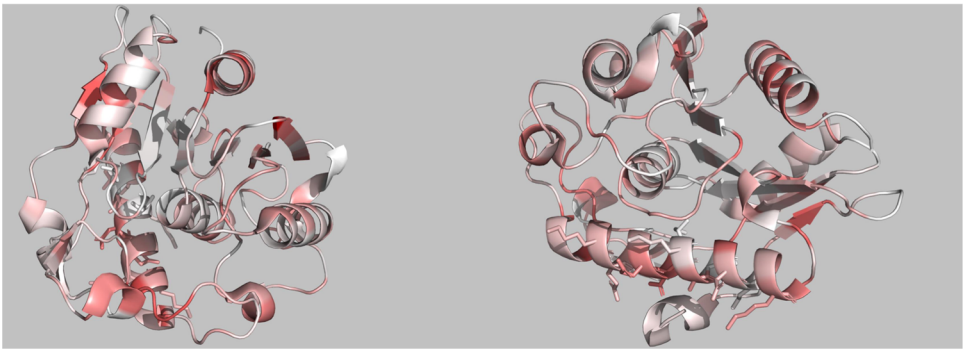


**Figure 4.** TE21 structure 1FJ2 with an $\alpha/\beta$ hydrolase fold. Darker red indicates residues with a higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein. Residue side chains as sticks are shown to highlight the difference between interior and exterior residues.

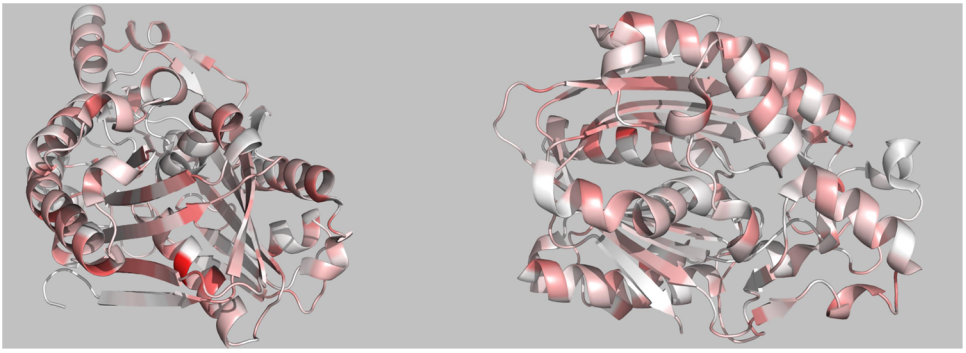


**Figure 5.** KS1 structure 4EFI. Darker red indicates residues with a higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein.

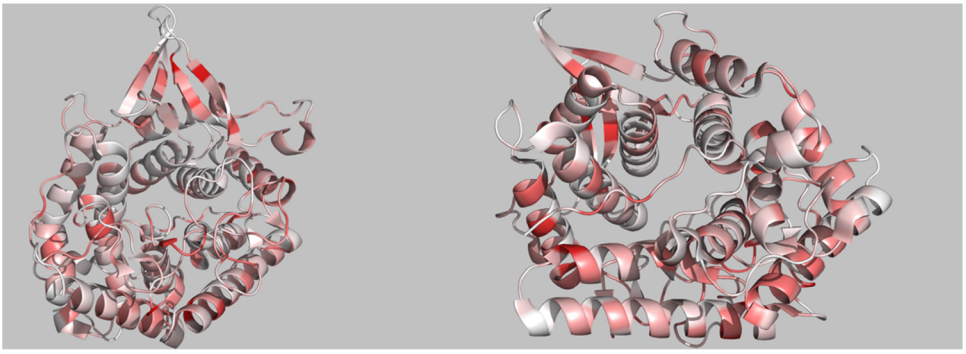


**Figure 6.** GH8 structure 1H12. Darker red indicates residues with a higher MSSC with more mutational space. Color scheme is normalized to the highest MSSC score in the protein.

The MSSC of each residue in the same target structure (1SC0) was recalculated, with an increased number of subject structures to include all of the structures in TE11 as well as those in TE6 (Figure 3b) to assess how the MSSC score is affected by including subject structures with less similarity with the target structure. TE6 enzymes also have a HotDog fold, but they have less structural similarity with each other than TE11 enzymes. Comparing Figure 3a,b, the regions with a higher mutation space are conserved; however, there are two main differences: (i) more residues have a higher mutation space as can be seen with more areas in light pink in Figure 3a,b and (ii) the residues with the higher MSSC scores changed, as seen, for example, by the sticks in the far left of both Figure 3a,b.

TE21[33] is also a thioesterase enzyme family, however, their proteins have an $\alpha/\beta$ hydrolase fold. The second example case is the twenty-one structures in TE21 with the *Homo sapiens* enzyme[35] (PDB 1FJ2) as the target structure (Figure 4). As in the TE11 example, secondary structures have higher MSSC scores (i.e., greater mutation space) in positions that are more solvent-exposed, as shown in the $\alpha$ helices present in 1FJ2. Residues side chains on an $\alpha$ helix have been represented as sticks to demonstrate the consistency of this phenomenon across secondary structures.

Ketoacyl synthase (KS) enzyme family KS1[36] has a thiolase-like fold and currently has forty-six resolved protein structures. The third example case is the 46 KS1 structures with the *Paraburkholderia xenovorans* structure[37] (PDB 4EFI) as the target protein (Figure 5). Of the four examples studied, KS1



https://doi.org/10.1021/acsomega.3c01473
*ACS Omega* 2023, 8, 24302−24310

enzymes show the most homogeneous mutation space with less variation between solvent-exposed and protein-exposed residues. Protein conformational changes and flexibility may play a role in the consistency with which some residues interact with one another, homogenizing the mutation space distribution. Further study may reveal that the homogeneity of mutation space distribution is fold-dependent.

GH8, a glycoside hydrolase enzyme family,[38] has structures with an $(\alpha/\alpha)_6$ barrel fold. The fourth example case is 44 GH8 structures, with a *Pseudoalteromonas haloplanktis* xylanase[39] (PDB 1H12) as the target protein (Figure 6). The mutation space of residues lying in closer proximity to other residues is generally lower than those that are more solvent-exposed. There is no significant difference in mutation space between secondary structures. It is worth noting that the "core" of this structure, likely the substrate-binding site, is a region of very low mutation space (very low MSSC scores). This suggests that residues involved with substrate specificity may also have low MSSC scores even if they are solvent-exposed.

For structure alignment comparison purposes, the mutation space of 1SC0 in TE11 was also calculated with structures superimposed with TM-Align and with Multiprot.[11,27] The overall structure superimposition of the TE11 structures is visually very similar (see Figure S1 in the Supporting Information). However, there are some differences in the MSSC scores: the TM-Align superimposed structures have the same MSSC score (+/−5) in 63% of 1SC0 residues as those aligned with our method, while the Multiprot aligned structures have the same MSSC score (+/−5) in 59% of 1SC0 residues as those aligned with our method. Five (5) is the smallest Grantham's distance that is between Leu and Ile. The difference in MSSC score comes from the differences in the structural alignment, where those performed by Multiprot or TM-Align result in optimized superimposed structures with the most possible spatial overlap (Figure S1). Although the MSSC score can be quantified for structures superimposed with other methods, the structure alignment method in this work (Section 2.1) was developed to identify pairs of spatially corresponding residues, not to optimize structure superimposition, locally or globally.

MSSC has a limitation since spatially corresponding residues are determined based on the similarity of their locations relative to common reference points ($\mathbf{CR}^1$,$\mathbf{CR}^2$, and **COM**). Therefore, the position of these reference points relative to each other and relative to their respective protein structures must be highly conserved. The example cases studied groups of protein structures (target and subject structures) within protein families, i.e., structures that have the same fold and a high degree of structural similarity; therefore, sequence-conserved residues are very highly conserved in their spatial positions, which gives consistent structural realignments and comparisons. If the set of proteins studied were more varied, the sequence-conserved residues would likely not be spatially conserved and the approach of identifying spatially conserved residues (Section 2.1) may not accurately determine spatially correlated residues. For each case in the Discussion section, Table S1 in the Supporting Information reports the total number of subject structures ($x$), average identities and similarities in target−subject pairwise sequence alignments, as well as the number of identities and similarities in the multiple sequence alignment of all of the sequences, which are used to select $\mathbf{CR}^1$ and $\mathbf{CR}^2$. The MSSC code can handle any set of structures as long as there are identities (*) and/or similarities (:) in the multiple sequence alignment that corresponds to spatially distant residues to select $\mathbf{CR}^1$ and $\mathbf{CR}^2$. Groups of protein structures whose average pairwise sequence alignments have at least ∼25% identity usually result in multiple sequence alignments with enough identities (*) and/or similarities to select $\mathbf{CR}^1$ and $\mathbf{CR}^2$.

The MSSC code is, for now, limited to monomeric protein chains. For multidomain proteins, the specific domain (monomer) of interest should be isolated and compared with similar structures.

Sequence and structure alignments do not always match, although a recently developed substitution matrix results in an improved agreement.[26] In our sample cases, some residues that are fully conserved in a multiple sequence alignment did not receive an MSSC of zero (i.e., fully conserved spatially as well), as one might expect. As an example, KS1 Gln274 is fully conserved in the multiple sequence alignment, however, its MSSC score in the *Paraburkholderia xenovorans* target protein is 1.69, which means that, in at least one other structure, a different residue occupies the spatial position that the target Gln274 residue does in its structure. MSSC focuses on the positions of residues relative to other residues and reference points within the structure. Therefore, minor variations in bond angles and residue size can shift residues from their expected spatial position, resulting in an MSSC that points to a different mutation space than expected. However, this is a natural consequence of our intent; we aim to examine and compare proteins through the space that their amino acid residues occupy, offering a different perspective than multiple sequence alignments or protein superimpositions. The function and stability of a protein are both intimately tied to its structure, and MSSC provides a new metric through which to compare and examine protein structures.

## 4. CONCLUSIONS

For each amino acid in a protein, the mutation space of spatially conserved amino acid sites (MSSC) quantifies how many, and how many different, mutations occur in a spatially defined amino acid site in a group of protein structures. Analysis of the mutation space of four protein structures in their respective families revealed that mutations are not uniform throughout the protein structure; rather, the composition of amino acid positions in a target structure varies in ordered ways. Regardless of the secondary structure, residue positions in closer proximity to other residues showed a lower MSSC score (i.e., more highly conserved), and those that lie further or are more solvent-exposed had a higher MSSC score (i.e., more frequently mutated), reproducing the well-known observation that solvent-exposed residues are more likely to be mutated than internal ones. These results demonstrate how quantifying the MSSC of residues in a protein can be used to understand the structural similarity of a protein more thoroughly within a set of related and structurally similar proteins.

## 5. METHODS

The Results section describes how spatially conserved amino acid sites are identified and how the MSSC is calculated for each residue in a protein. Data processing was performed with Python 3.8. Custom scripts were written for all data processing and the NumPy math package was utilized for nontrivial math functions. Throughout this work, the spatial position of


https://doi.org/10.1021/acsomega.3c01473
ACS Omega 2023, 8, 24302−24310

residues within a protein structure is defined by the cartesian coordinates of reference vectors ending at the $\alpha$-carbon atom of that residue. Multiple sequence alignments were performed using MUSCLE[40] using the default settings and output for processing in the CLUSTALW format.

The MSSC code is available on GitHub: "https://github.com/Cantu-Research-Group/Mutation_Space". See README file for specific instructions.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c01473.

Target and subject PDB structures in the Discussion section; case 1 superimposed with three different structure alignment methods; Figure S1: Different structure superimposition methods; pairwise sequence alignment identities and similarities; Table S1: MSSC performance and sequence similarity; MSSC code (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**David C. Cantu** − *Department of Chemical and Materials Engineering, University of Nevada, Reno, Nevada 89557, United States;* orcid.org/0000-0001-9584-5062; Email: dcantu@unr.edu

### Authors

**Benjamin T. Caswell** − *Department of Chemical and Materials Engineering, University of Nevada, Reno, Nevada 89557, United States*

**Thomas J. Summers** − *Department of Chemical and Materials Engineering, University of Nevada, Reno, Nevada 89557, United States;* orcid.org/0000-0002-4243-6078

**Gerra L. Licup** − *Department of Chemical and Materials Engineering, University of Nevada, Reno, Nevada 89557, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c01473

### Author Contributions

†B.T.C. and T.J.S. contributed equally to this work.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was supported by the U.S. National Science Foundation (Award 2001385).

## ■ REFERENCES


(1) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(2) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(3) Ferrer-Costa, C.; Gelpí, J. L.; Zamakola, L.; Parraga, I.; de la Cruz, X.; Orozco, M. PMUT: A Web-Based Tool for the Annotation of Pathological Mutations on Proteins. *Bioinformatics* **2005**, *21*, 3176−3178.

(4) Wang, Z.; Moult, J. SNPs, Protein Structure, and Disease. *Hum. Mutat.* **2001**, *17*, 263−270.

(5) Lee, W.; Yue, P.; Zhang, Z. Analytical Methods for Inferring Functional Effects of Single Base Pair Substitutions in Human Cancers. *Hum. Genet.* **2009**, *126*, 481−498.

(6) Yue, P.; Li, Z.; Moult, J. Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. *J. Mol. Biol.* **2005**, *353*, 459−473.

(7) Ng, P. C.; Henikoff, S. Predicting the Effects of Amino Acid Substitutions on Protein Function. *Annu. Rev. Genomics Hum. Genet.* **2006**, *7*, 61−80.

(8) Krebs, F. S.; Zoete, V.; Trottet, M.; Pouchon, T.; Bovigny, C.; Michielin, O. Swiss-PO: A New Tool to Analyze the Impact of Mutations on Protein Three-Dimensional Structures for Precision Oncology. *npj Precis. Oncol.* **2021**, *5*, No. 19.

(9) Chothia, C.; Lesk, A. M. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5*, 823−826.

(10) Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Šali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.

(11) Shatsky, M.; Nussinov, R.; Wolfson, H. J. A Method for Simultaneous Alignment of Multiple Protein Structures. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 143−156.

(12) Holm, L. DALI and the Persistence of Protein Shape. *Protein Sci.* **2020**, *29*, 128−140.

(13) Ortiz, A. R.; Strauss, C. E. M.; Olmea, O. MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison. *Protein Sci.* **2009**, *11*, 2606−2621.

(14) Heidarian, A.; Dinneen, M. J. A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering. *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, Oxford, UK, 2016; pp 142−151.

(15) Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **1974**, *185*, 862−864.

(16) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.

(17) Miyazawa, S.; Jernigan, R. L. A New Substitution Matrix for Protein Sequence Searches Based on Contact Frequencies in Protein Structures. *Protein Eng., Des. Sel.* **1993**, *6*, 267−278.

(18) Jung, J.; Lee, B. Use of Residue Pairs in Protein Sequence−Sequence and Sequence−Structure Alignments. *Protein Sci.* **2000**, *9*, 1576−1588.

(19) Yu, Y.-K.; Wootton, J. C.; Altschul, S. F. The Compositional Adjustment of Amino Acid Substitution Matrices. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15688−15693.

(20) Vilim, R. B.; Cunningham, R. M.; Lu, B.; Kheradpour, P.; Stevens, F. J. Fold-Specific Substitution Matrices for Protein Classification. *Bioinformatics* **2004**, *20*, 847−853.

(21) Liu, X.; Zhao, Y.-P. Substitution Matrices of Residue Triplets Derived from Protein Blocks. *J. Comput. Biol.* **2010**, *17*, 1679−1687.

(22) Yamada, K.; Tomii, K. Revisiting Amino Acid Substitution Matrices for Identifying Distantly Related Proteins. *Bioinformatics* **2014**, *30*, 317−325.

(23) Tomii, K.; Yamada, K. Systematic Exploration of an Efficient Amino Acid Substitution Matrix: MIQS BT - Data Mining Techniques for the Life Sciences. In *Methods in Molecular Biology*; Carugo, O.; Eisenhaber, F., Eds.; Springer New York: New York, NY, 2016; pp 211−223.

(24) Keul, F.; Hess, M.; Goesele, M.; Hamacher, K. PFASUM: A Substitution Matrix from Pfam Structural Alignments. *BMC Bioinf.* **2017**, *18*, No. 293.

(25) Edgar, R. C. Optimizing Substitution Matrix Choice and Gap Parameters for Sequence Alignment. *BMC Bioinf.* **2009**, *10*, No. 396.

(26) Jia, K.; Jernigan, R. L. New Amino Acid Substitution Matrix Brings Sequence Alignments into Agreement with Structure Matches. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 671−682.

(27) Zhang, Z.; Wang, Y.; Wang, L.; Gao, P. The Combined Effects of Amino Acid Substitutions and Indels on the Evolution of Structure within Protein Families. *PLoS One* **2010**, *5*, No. e14316.

(28) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(29) Matthews, B. W. Studies on Protein Stability With T4 Lysozyme. In *Advances in Protein Chemistry*; Academic Press, 1995; Vol. *46*, pp 249−278.
(30) Cordes, M. H. J.; Sauer, R. T. Tolerance of a Protein to Multiple Polar-to-Hydrophobic Surface Substitutions. *Protein Sci.* **2008**, *8*, 318−325.
(31) Hill, R. B.; DeGrado, W. F. A Polar, Solvent-Exposed Residue Can Be Essential for Native Protein Structure. *Structure* **2000**, *8*, 471−479.
(32) van den Burg, B.; Eijsink, V. G. H. Selection of Mutations for Increased Protein Stability. *Curr. Opin. Biotechnol.* **2002**, *13*, 333−337.
(33) Caswell, B. T.; de Carvalho, C. C.; Nguyen, H.; Roy, M.; Nguyen, T.; Cantu, D. C. Thioesterase Enzyme Families: Functions, Structures, and Mechanisms. *Protein Sci.* **2022**, *31*, 652−676.
(34) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(35) Devedjiev, Y.; Dauter, Z.; Kuznetsov, S. R.; Jones, T. L. Z.; Derewenda, Z. S. Crystal Structure of the Human Acyl Protein Thioesterase I from a Single X-Ray Data Set to 1.5 Å. *Structure* **2000**, *8*, 1137−1146.
(36) Chen, Y.; Kelly, E. E.; Masluk, R. P.; Nelson, C. L.; Cantu, D. C.; Reilly, P. J. Structural Classification and Properties of Ketoacyl Synthases. *Protein Sci.* **2011**, *20*, 1659−1667.
(37) Baugh, L.; Gallagher, L. A.; Patrapuvich, R.; Clifton, M. C.; Gardberg, A. S.; Edwards, T. E.; Armour, B.; Begley, D. W.; Dieterich, S. H.; Dranow, D. M.; Abendroth, J.; Fairman, J. W.; Fox, D., III; Staker, B. L.; Phan, I.; Gillespie, A.; Choi, R.; Nakazawa-Hewitt, S.; Nguyen, M. T.; Napuli, A.; Barrett, L.; Buchko, G. W.; Stacy, R.; Myler, P. J.; Stewart, L. J.; Manoil, C.; Van Voorhis, W. C. Combining Functional and Structural Genomics to Sample the Essential Burkholderia Structome. *PLoS One* **2013**, *8*, No. e53851.
(38) Lombard, V.; Ramulu, H. G.; Drula, E.; Coutinho, P. M.; Henrissat, B. The Carbohydrate-Active Enzymes Database (CAZy) in 2013. *Nucleic Acids Res.* **2014**, *42*, D490−D495.
(39) Van Petegem, F.; Collins, T.; Meuwis, M.-A.; Gerday, C.; Feller, G.; Van Beeumen, J. The Structure of a Cold-Adapted Family 8 Xylanase at 1.3 Å Resolution: Structural Adaptations to Cold and Investigation of the Active Site. *J. Biol. Chem.* **2003**, *278*, 7531−7539.
(40) Edgar, R. C. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res.* **2004**, *32*, 1792−1797.


https://doi.org/10.1021/acsomega.3c01473
*ACS Omega* 2023, 8, 24302−24310