ClinicSearch

International Journal of Clinical Surgery

Aminul Islam *

Open Access

Research Article

Protein secondary structure and remote homology detection

Ali Al-Fatlawi [1,3], Md. Ballal Hossen [1], Ferras El-Hendi [1,2] and Michael Schroeder [1,2]*

[1]Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering, Technische Universit¨at Dresden, Dresden, Germany.

[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden, Germany.

[3]University of Kufa, Iraq.

Corresponding Author: Michael Schroeder. Biotechnology Center (BIOTEC), Center for Molecular and Cellular Bioengineering, Technische Universit¨ and Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden, Germany.

Received Date: May 03, 2024; Accepted Date: May 13, 2024; Published Date: May 27, 2024

Citation: Ali Al-Fatlawi, Md. Ballal Hossen, Ferras El-Hendi and Michael Schroeder. (2024), Protein secondary structure and remote homology detection, *International Journal of Clinical Surgery,* 3(3); **DOI:10.31579/2834-5118/049**

Abstract

A protein can be represented by its primary, secondary, or tertiary structure. With recent advances in AI, there is now as much tertiary as primary structural data available. Fast and accurate search methods exist for both types of data, with searches over both representations being highly precise. However, primary structure data can sometimes be incomplete. As a result, tertiary structure has become the gold standard for remote homology detection.

How does secondary structure perform in remote homology detection? Secondary structure interprets proteins as a sequence using an alphabet representing helices, strands, or loops. It shares its sequential nature with primary structure while retaining topological information similar to tertiary structure.

To assess the effectiveness of secondary structure in remote homology detection, we devised a challenging classification task aimed at determining the superfamily membership of very distantly related protein domains. We used benchmarks from the CATH and SCOP databases and evaluated sequence and structure alignment algorithms on primary, secondary, and tertiary structures. As expected, both basic and advanced sequence alignment algorithms applied to primary structure achieved high precision, but their overall area under the curve was lower compared to the gold standard of structural alignment using tertiary structure.

Surprisingly, a simple string comparison algorithm applied to secondary structure performed close to the gold standard. This result supports the hypothesis that key structural information is already encoded in secondary structure and suggests that secondary structure may be a promising representation to use when high-confidence structural data is unavailable, such as in cases involving protein flexibility and disorder.

Keywords: homology detection; protein flexibility and disorder

Introduction

The primary structure represents proteins as sequences of amino acids, while the tertiary structure provides a set of atomic coordinates of these amino acids, which form helices, strands, or loops in 3D space. The secondary structure serves as an intermediate representation, capturing helices, strands, and loops in a sequential format. Generally, tertiary structure is more conserved than pri- mary structure, as functional requirements impose constraints on a protein's structure, whereas the sequence itself can mutate as long as essential functions are preserved [1]. As a result, dissimilar sequences may fold into similar struc- tures that are homologous in function [2, 3]. Functional constraints often apply only to specific regions of the protein rather than the entire structure, and a substantial portion of the protein sequence can undergo significant variation as long as the core structure, which satisfies functional requirements, remains intact [4]. Consequently, primary structure alignment tools like BLAST [5] and HHblits [6] are highly precise in remote homology detection, but they are not complete for the reasons mentioned above. On the other hand, tertiary structure serves as a gold standard, offering both precision and completeness [3].

In contrast to sequence alignments, naive structural alignments performed by tools like TM-align [7] and CE-align [8] are computationally intensive, as these algorithms determine similarity by finding an optimal superposition of atomic coordinates. Such methods are not suitable for searching large structural repositories, such as the AlphaFold database [9] or the ESM

Metagenomic Atlas [10], which contain hundreds of millions of structures. Therefore, 3D fins- gerprints have been developed and implemented in the structure search engine Foldseek [11]. These 3D fingerprints represent proteins as sequences over an abstract alphabet, capturing local information about the atoms' coordinates and interactions. In summary, fast and accurate searches over primary and tertiary structures are available with BLAST and FoldSeek, respectively. The former is highly precise in detecting remote homologs, while the latter is both precise and complete.

How does secondary structure perform in homology detection? Does it share the precision of BLAST but fail to retrieve all remote homologs, or does its abstract representation hold sufficient structural information to perform as well as tertiary structure? The latter hypothesis is supported by Przytycka et al., who explored the extent to which a protein's secondary structure could inform its three-dimensional fold by analyzing known protein structures [12]. They constructed a taxonomy based solely on secondary structure, proposing a simple mechanism of protein evolution [12]. Fontana et al. and Guharoy et al. found that secondary structure is sufficiently conserved to compute align- ments of protein secondary structures against a library of domain folds [13] and to identify binding motifs in protein-protein interactions, respectively [14]. A concrete example of the conservation of secondary structure is the superfam- ily of single-strand annealing proteins, which comprises five distantly related families, all sharing a secondary structure motif of a β-hairpin flanked by two helices and a β-sheet with a perpendicular helix [15] (see Figure 1).

In addition to the conservation of secondary structure, a second reason for its effectiveness is that very fast sequence alignment algorithms can be applied to secondary structure, a key factor in the fast structure search capabilities of FoldSeek using 3D fingerprints.

To answer the question of how secondary structure compares to primary and tertiary structures in remote homology detection, we focus on a challenging task: determining superfamily membership for a non-redundant set of structural domains from the CATH and SCOP domain databases. Non-redundancy refers to sequence similarity, making this task naturally difficult for sequence algorithms based on primary structure. For all three representations, we mea- sure the similarity of domain pairs from the same superfamily versus those from different superfamilies. We aim to determine whether the performance of alignments based on secondary structure is closer to the gold standard of tertiary structure or to the highly precise but incomplete primary structure.

2. Results and Discussion

Secondary structure alphabet

Secondary structure is an abstraction of 3D structure. In 3D, the connection between two consecutive amino acids is described by two angles, known as the phi and psi angles. These angles cannot adopt all theoretical combinations but instead cluster around certain values. These clusters lead to the assignment of secondary structure to a residue. Depending on the clustering, the secondary structure can be represented using a more restrictive or a less restrictive alpha- bet. Typically, two representations are used: a 3-letter alphabet (helix, strand, loop) and an 8-letter alphabet (which includes three types of helices 310-helix, α-, and π-helix—strand, loop, and three additional letters for specialized turns, strands, and coils). We compare both representations to determine whether the increased granularity of the 8-letter alphabet leads to improved performance.

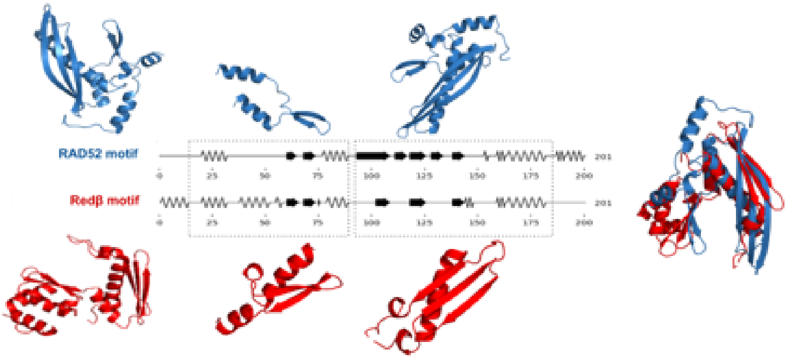


Figure. 1: Diagram illustrating the secondary sequence alignment between Rad52 and Redβ as an example from SSAPs.

	CATH	SCOPe40	CATH S20
Redundancy	None	40%	20%
Domains	23,911	11,160	14,907
Pairs	285,856,005	62,267,220	111,101,871
Superfamilies	1,075	1,954	3,288
Pairs in the same Superfamily	6,348,666 (2.2%)	225,931 (0.36%)	582,578 (0.52%)
Pairs in different Superfamilies	279,507,339	62,041,289	110,519,293

Benchmark datasets

To compare primary, secondary, and tertiary structures, we employed a chal- lenging classification task. We compared scores within predefined domain superfamilies and between distinct superfamilies, where the former should yield better scores. The two principal databases for structural domain classification are CATH [16] and SCOPe [17], which organize protein domains by topologyat the family and superfamily levels. We included both databases and varied the degree of redundancy in our datasets. A non-redundant dataset is more challenging, especially for sequence-based methods, but also more realisticfor scenarios involving remote homology detection, where little prior knowl- edge exists. Therefore, we devised three datasets: CATH (with redundancy), SCOPe40 (non-redundant at 40% sequence similarity), and CATH S20 (non- redundant at 20%). Each dataset covers over 1,000 superfamilies and more than 10,000 individual domains. The number of pairs ranges from 62 to 286 million, with the number of pairs within the same superfamily being a small fraction of those between different superfamilies (see Table 1), ranging from 0.36% to 2.2%. This classification task is highly unbalanced, reflecting the task of remote homology detection, where the vast majority of relationships are negative. The ratio of positive to all pairs in the two non-redundant datasets, SCOPe40 and CATH S20, is an order of magnitude smaller than in the redun- dant dataset, CATH, highlighting the increased difficulty of these datasets for the algorithms.

Algorithms.

The Levenshtein distance [18] is the most basic sequence comparison algorithm, computing the minimal number of insertions or deletions necessary to con- vert one sequence into another. For comparing secondary structure sequences, we used the Levenshtein distance with one modification: we normalized it to account for substantial variations in sequence length (see Methods). The Lev- enshtein distance also forms the basis for advanced algorithms like BLAST, which uses an enhanced scoring scheme for gaps and mismatches and is opti- mized for speed. While BLAST is the standard method for amino acid sequence comparison, specialized approaches optimized for remote homology detection exist, such as those using hidden Markov models (HMMs). HMMs generate a statistical representation of a protein family, which is more robust than individ- ual sequences. A widely used HMM implementation is HHblits [6]. For tertiary structure, the most widely used comparison method translates 3D structures into 3D fingerprints, sequences that can be interpreted as high-dimensional vectors, allowing for very efficient comparison methods. An example of such an approach is Foldseek [19]. However, since these methods aim to approximate the slower, optimal superposition of atomic coordinates, we used TM-align [20] as a reference and gold standard.

In summary, we compared primary structure using basic BLAST and advanced HHblits methods, secondary structure using a normalized Leven- shtein distance with 3-letter and 8-letter alphabets, and tertiary structure using TM-align. These were applied to the three benchmark datasets: CATH, SCOPe40, and CATH S20.

Secondary structure's performance is closer to tertiary's than to primary's

Primary, secondary, and tertiary structures achieved AUCs of up to 84%, 95%, and 98%, respectively, across the three benchmarks and varying setups (see Table 2 and Figure 2). This indicates that a basic sequence alignment algorithm, such as the Levenshtein distance applied to secondary structure, significantly outperforms both basic and advanced algorithms on primary structure and approaches the performance of tertiary structure. This sug- gests that the topological information embedded in secondary structure can be effectively utilized, even with simple alignment algorithms.

Non-redundant datasets are consistently more challenging

The more non-redundant a dataset is, the fewer "easy" remote homologues exist, making the dataset more challenging. Surprisingly, this is true not only for algorithms that explicitly leverage redundancy, such as HHblits, but also for secondary and tertiary structure-based methods.

Performance of HMMs increases with the amount of sequences

HMMs capture a statistical signature of a sequence family, making them more effective for remote homology detection than basic sequence comparison. As a result, HMMs consistently outperform BLAST, especially as more sequence data becomes available. For instance, HMMs achieved 84% AUC on the CATH dataset, compared to 51% for BLAST on the highly redundant CATH S20. This demonstrates the effectiveness of HMMs in leveraging large sequence datasets. The recent success of embeddings computed from large language models for amino acid sequences [21, 22] builds on this effect.

TM-Score is a gold standard

Structural alignment of tertiary structures across the full CATH dataset nearly perfectly classifies (98%) domain pairs as remote homologues or not. However, its performance drops by 8% with reduced redundancy. One reason for this is that remote homologues can vary in structure, and structural flexibility can affect alignment accuracy.

The Secondary structure alphabet does not affect classification

One might assume that a more detailed representation of secondary structure would lead to improved results, but this is not the case. The AUC results are nearly identical for both the 3-letter and 8-letter representations. This is because, unlike amino acid alphabets, where large differences in frequencies can be attributed to physicochemical properties (e.g., the rarity of cysteines form- ing disulfide bonds), such specific roles are not ascribed to the more detailed secondary structure descriptions of the 8-letter alphabet (see Supplementary Material, Note 1).

Secondary structure performs nearly as well as the gold standard

On the CATH dataset, secondary structure achieved an AUC only 3% lower than the tertiary structure gold standard. For the most difficult benchmark, this margin increases to 9%. However, compared to HMMs, which perform 32% worse than the gold standard, this is a remarkable result.

Substitution matrices with local alignment have potential

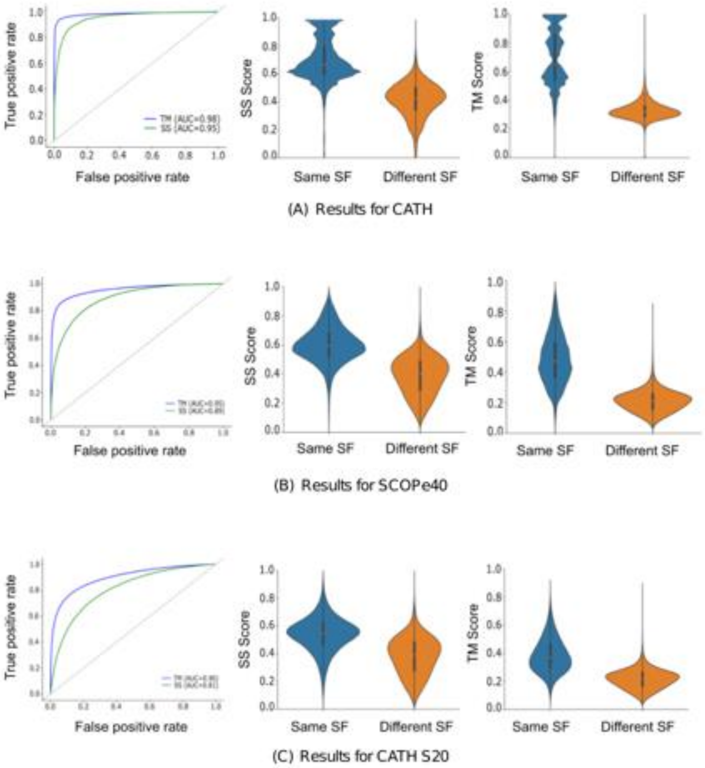
To explore whether the 9% gap can be further reduced, we turned to more advanced sequence comparison algorithms. The Levenshtein algorithm does not employ scoring with varying gap penalties or consider likely versus unlikely mismatches. However, even though secondary structure lacks the fine granu- larity of primary structure, differences exist—for example, helix residues are more likely to be replaced by loop residues than by strand residues. This was observed in [23], where a scoring scheme for secondary structure was developed. To this end, we created two substitution matrices for secondary structure, inspired by BLOSUM [24]. Using high-quality multiple sequence alignments from PFAM [25], we mapped the amino acids to the corresponding secondary structure letters and computed substitution frequencies. The result- ing secondary structure substitution matrix was used in a local alignment with the Smith-Waterman algorithm. On the challenging CATH S20 dataset, this approach improved performance by 5%, from 80% to 85%, narrowing the gap to the gold standard from 9% to 5% (see Supplementary Material, note 2 and note 3).

Structure	Methods	CATH	SCOPe40	CATH S20
Primary (amino acid)	BLAST	0.66	0.56	0.51
	HHblits	0.84	0.77	0.58
Secondary (SS string)	Levenshtein with 3-letter	0.95	0.89	0.81
	Levenshtein with 8-letter	0.95	0.90	0.81
Tertiary (3D)	TM-score	0.98	0.95	0.90

Table 2: Performance (AUCs) comparison of different methods on CATH, SCOPe40, and CATH S20 datasets.

Setting a threshold for secondary structure

In this study, we focused on comparing representations for which AUC is an adequate measure. However, to make practical use of secondary structure in homology detection, it is crucial to assess the likelihood of a score and establish a threshold. We utilized Bayes' Theorem to calculate the posterior probabilities of protein pairs belonging to the same superfamily based on their secondary structure alignment scores, providing a probabilistic framework for interpret- ing these scores within the context of our dataset [26] (see Methods). The groups exhibited different secondary structure score ranges, with some over- lap between 0.5 and 0.7. Protein pairs within the same superfamily generally had higher alignment scores, predominantly within the 0.6 to 1.0 range. For further details, see Supplementary Material, Note 4.



(A) Results for CATH

(B) Results for SCOPe40

(C) Results for CATH S20

Figure 2: This figure presents the true positive rate (TPR) against the false pos- itive rate (FPR), along with their corresponding AUC values, for (A) CATH, (B) SCOPe40, and (C) CATH S20. The AUC plot compares the performance of tertiary structure alignment with TM-align and secondary structure with Levenshtein using sequences in its 3-letter representation." SS scores" refer to secondary structure alignment scores, and" SF" pertains to the superfamily. The violin plots show how the secondary structure approximates the informa- tion of the tertiary structure in separating proteins that belong to the same superfamily (Same SF) and those in different superfamilies (different SF).

3. Methods

Data collection

CATH version 4.3.0 and SCOPe version 2.07 were used. To maintain a fair com- parison, we kept proteins with lengths between 50 and 250, representing most of the domains. To narrow our dataset further, we only considered domains with a single selection range and had the correct selection range of residues in CATH with superfamily classification. This resulted in a benchmark of 23,911 domains. The non-redundant sets of CATH S20 and SCOPe40 were fully con- sidered without filtering; see Table 1. Secondary structures for these domains were extracted using Pymol (v 2.2.0 Open-Source) for the Cα atoms only.

Substitution matrix for secondary structure:

For the development of the substitution matrix, Pfam-A Seed alignments were used (downloaded 23/04/24) [25]. These were filtered to only include sequences for which AlphaFold structures existed and the alignment ranges matched. This resulted in 19,226 alignments of a total of 1,155,996 sequences. Secondary structures for these sequences were extracted again as above. In accordance with the original BLOSUM authors [24], the alignments were then trimmed to remove columns containing gaps, leaving 2,460,186 ungapped columns.

With the data prepared, we apply the same steps as for BLOSUM. Firstly, pairwise frequencies fij are counted for all columns and all pairs ij. Next, the observed probability qij is calculated as the frequency of a pair ij, relative to the number of all pairs:

4.Conclusion

The representation of proteins is an intriguing and open-ended question, with the answer depending on the specific purpose and requirements of the represen- tation. Until recently, there was a trade-off: representing proteins as primary structures allowed for a wealth of data and fast algorithms, which was not the case for tertiary structures. However, with the advent of AlphaFold and related systems, there is now a comparable amount of structural data, and with the development of FoldSeek, structural representations are also amenable to fast structural searches. As a result, fast and accurate remote homology detection is now possible, leveraging tertiary rather than primary structures.

While this question seems practically settled, one aspect remains open: how does secondary structure perform in remote homology detection? Secondary structure shares the sequential nature of primary structure and the topological information of tertiary structure. In this work, we address this question and report the surprising result that simple sequence comparison of three-letter secondary structure performs nearly as well as tertiary structure in distin- guishing domains from the same superfamily. Moreover, secondary structure offers an advantage that tertiary structure lacks: flexibility in tertiary struc- ture can make similarity detection difficult for structure alignment algorithms, whereas secondary structure is robust against such variations. Therefore, sec- ondary structure is a promising representation that should be considered when dealing with 3D protein structures.

5. Data Availability

The secondary structure strings, SS scores, TM scores, and superfamily memberships are available at the following link:

https://sharing.biotec.tu-dresden.de/index.php/s/eKaamoJtTJPLbJk

6.Acknowledgement

We gratefully acknowledge the financial support provided by the BMBF projects scads.ai and SNRT, as well as the access to high-performance com- puting resources through the ZIH of TU Dresden, which were instrumental in the completion of this study. We extend our sincere thanks to Francis Stew- art and Christian Pilarsky for their valuable discussions and insights, which greatly contributed to the development of this research.

7.Author contributions

AAF and MS conceived the study. AAF, BH, FE, and MS analyzed data and wrote the manuscript.

8.Competing Interests: The Authors declare no competing interests.

Supplementary Material

Supplementary Note 1:

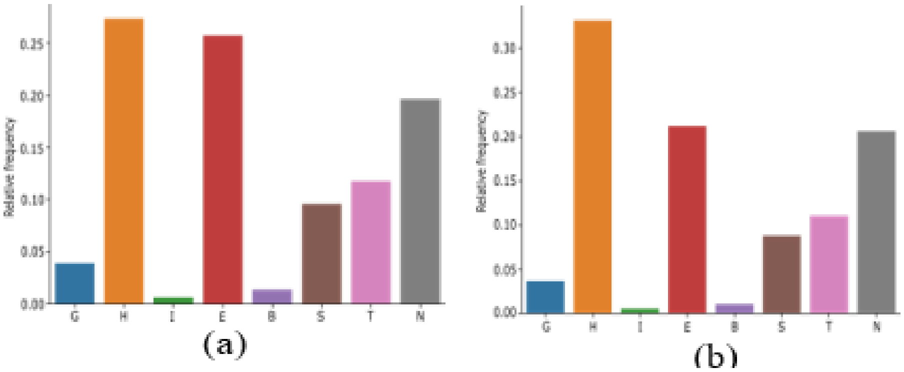


Figure. S1: The frequency of each letter in the 8-letter representation secondarystructure according to the DSSP method. (a) CATH dataset (b) SCOPe40

Supplementary Note 2:

The likelihood of replacement for three secondary structure letters (S, H, L) in CATH and SCOPe40

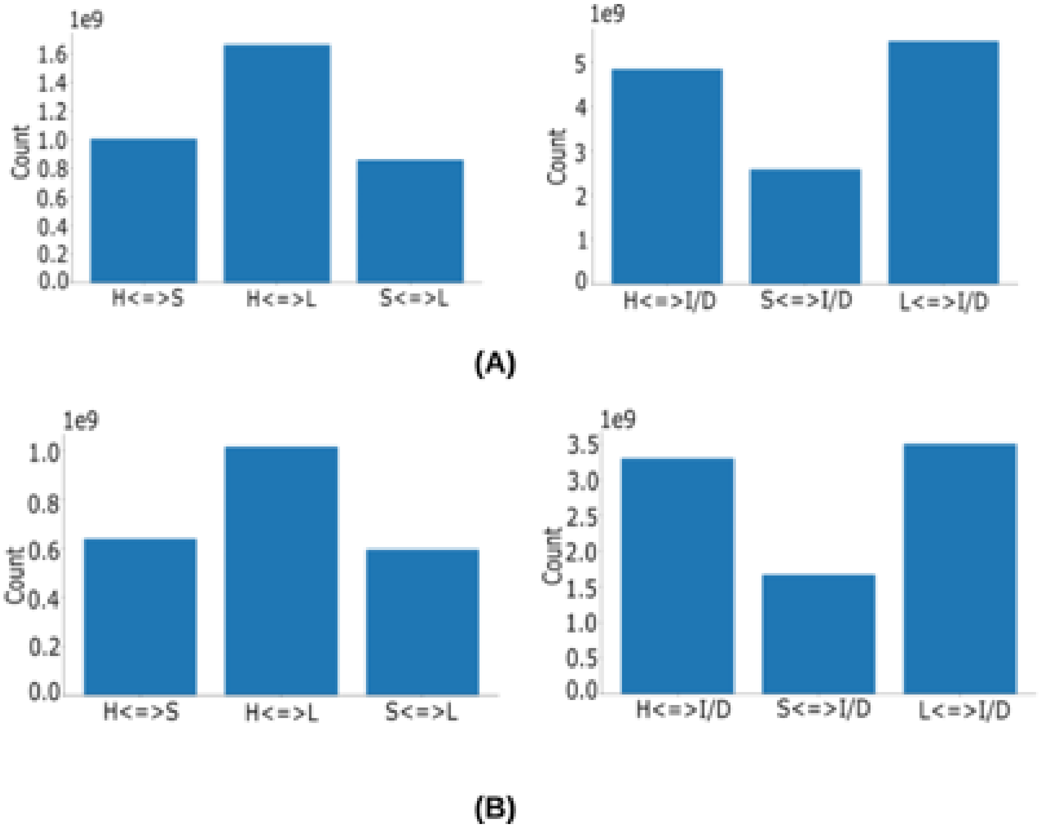


Figure. S2: Substitution counts for letters in CATH (A) and SCOPe40 (B) datasets. 'H' represents helices, 'S' represents sheets, 'L' represents loops, and 'I/D' stands for insertion/deletion events. The ¡=¿ symbol indicates bidi- rectional substitution. In both panels (A) and (B), the left

figures show substitution counts between ’H’, ’S’, and ’L’, while the right figures display substitution counts among ’H’, ’S’, ’L’, and insertions/deletions.

Supplementary Note 3:

Custom substitution matrix for secondary structure derived using log-odds ratios from Pfam-A Seed alignments

	H	S	L
H	2		
S	-7	3	
L	-16	-5	4

Table S1: Pairwise substitution values used in the Smith-Waterman alignment algorithm for SS strings.

(a)

	H	S	L
H	1		
S	-7	4	
L	-16	-3	4

(b)

	H	S	L
H	2		
S	-8	3	
L	-16	-5	4

(c)

	H	S	L
H	2		
S	-7	3	
L	-16	-5	4

(d)

	H	S	L
H	2		
S	-7	3	
L	-16	-5	4

Table S2: In order to verify the robustness of the matrix, we re-calculated the matrix for smaller sample sizes, of a) 500, b) 1,000, c) 5,000 and d) 10,000 alignments, that were randomly sampled from the full Pfam-A Seed alignment set. As can be seen from a) and b) for smaller sample sizes, there are minor deviations of magnitude up to 2 in a) and 1 in b) respectively, however for a sample size of 5,000 the substitution matrix values match the default matrix’s values, such that it is safe to assume our matrix to be robust. We ran a similar experiment to test robustness against sequence similarity, where matrices were computed using only alignments of maximum pairwise similarity of 20%, 45%, 50%, 62%, 80%, and 90% respectively. All of these matched the original matrix, which is not surprising since domains are considered to have structurally conserved motifs

.

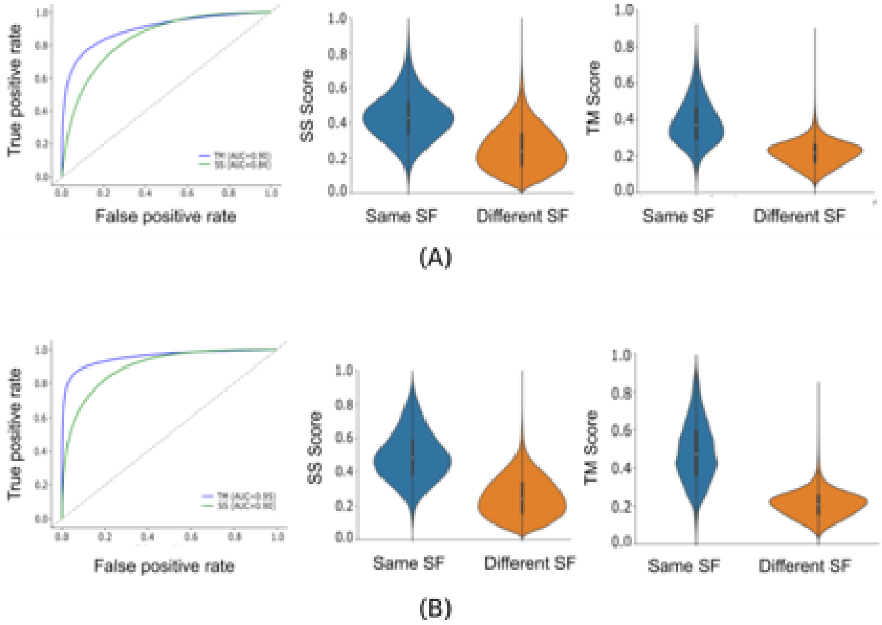


Figure. S3: True Positive Rate (TPR) vs. False Positive Rate (FPR) and cor- responding AUC values for CATH S20 (A) and SCOPe40 (B), using a local alignment with a customized substitution matrix. We set the penalty for replacing ’S’ with ’L’ or ’H’ with ’L’ (i.e., -0.5) lower than for replacing ’S’ with ’H’ or vice versa, which was set at -1 (see Methods). This customization was driven by the observed distribution of replacements between these letters. Here, we used an alternative Needleman-Wunsch algorithm [15] for local align- ment. These modifications improved performance in CATH from 81% to 84% and slightly enhanced performance in SCOPe from 89% to 90%.

Supplementary Note 4:

Further Statistical Analysis

We calculate the conditional probability for a given secondary structure align- ment score (SS score) of protein pairs being in the same superfamily. For an SS score below 0.6, the probability of a protein pair belonging to the same superfamily is close to 0; this is the case for only a few pairs. However, we can also observe that for an SS score above 0.8, that probability increases sharply to around 50%. Around the score of 0.85, we observe a clear phase transi- tion. There are also a few unexpected spikes or respective drops, especially for SS scores around 1.0, but these are rather outliers and can be explained by anomalies in and the reduced size of the dataset. Nonetheless, this phase transition points at a possible threshold of 0.85, indicating that two proteins can be expected to be in the same superfamily or superfamilies. The same procedure is done on the TM score in our data, and as was underlined in previous studies [31], the threshold of 0.5 is optimal for superfamily classification, figure below. We also performed the same analysis on the SCOPe dataset and for several methods, i.e., TM score and SS score using a 3-letter secondary struc- ture assignment, consisting of a three-letter alphabet, and the DSSP secondary structure assignment, consisting of an eight-letter alphabet [32, 33].

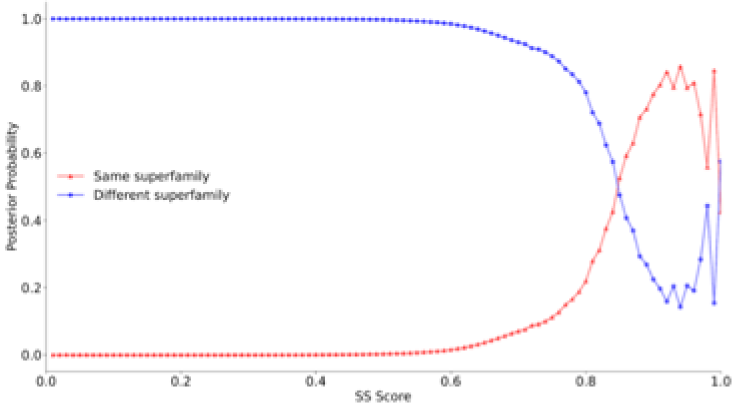


Figure S4: The posterior probabilities of SCOPe domain pairs for a given secondary structure score being in the same superfamily (red) or different superfamilies (blue). Both lines cross at around 0.85.

References:

1. Lesk, A.M. (2000). Introduction to Protein Architecture: The Structural Biology of Proteins. *Oxford University Press*.
2. Krissinel, E. (2007). On the relationship between sequence and structure simi- larities in proteomics. *Bioinformatics* 23(6), 717–723.
3. Al-Fatlawi, A., Menzel, M., Schroeder, M. (2023). Is protein blast a thing of the past. *Nat Commun* 14(1), 8195.
4. Schulz, G.E., Schirmer, R.H. (2013). Principles of Protein Structure. *Springer*.
5. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215(3), 403–410.
6. Remmert, M., Biegert, A., Hauser, A., S¨oding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 9(2), 173–175
7. Xu, J., Zhang, Y. (2010). How significant is a protein structure similarity with tm-score = 0.5. *Bioinformatics* 26(7), 889–895.
8. Shindyalov, I.N., Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11(9), 739–747
9. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596(7873), 583–589.
10. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., et al., (2023). Evolutionary-scale prediction of atomic- level protein structure with a language model. *Science* 379(6637), 1123– 1130.
11. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., et al., (2023). Fast and accurate pro-tein structure search with Foldseek. *Nature Biotechnology*.
12. Przytycka, T., Aurora, R., Rose, G.D. (1999). A protein taxonomy based on secondary structure. *Nat Struct Biol* 6(7), 672–682.
13. Fontana, P., Bindewald, E., Toppo, S., Velasco, R., Valle, G., Tosatto, S.C. (2005). The SSEA server for protein secondary structure alignment. *Bioinformatics* 21(3), 393–395.
14. Guha Roy, M., Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics* 23(15), 1909–1918
15. Al-Fatlawi, A., Schroeder, M., Stewart, A.F. (2023). The rad52 SSAP super- family and new insight into homologous recombination. *Communications Biology* 6(1).
16. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., et al., (1997). CATH–a hierarchic classification of protein domain structures. *Structure* 5(8), 1093–1108.
17. Fox, N.K., Brenner, S.E., Chandonia, J.-M. (2014). Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research* 42(D1), 304–309.
18. Levenshtein, V.I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet Physics Doklady*, vol. 10, pp. 707–710. Soviet Union
19. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Gilchrist, C.L.M., et al., (2022). Foldseek: fast and accurate protein structure search.
20. Zhang, Y., Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* 33(7), 2302–2309.
21. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., et al. (2021). Prottrans: Toward understanding the language of life through self-supervised learn- ing. IEEE transactions on pattern analysis and machine intelligence 44(10), 7112–7127
22. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., et al., (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373(6557), 871–876
23. Fischel-Ghodsian, F., Mathiowitz, G., Smith, T.F. (1990). Alignment of protein sequences using secondary structure: a modified dynamic programmingmethod. Protein Engineering, *Design and Selection* 3(7), 577–581.
24. Henikoff, S., Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. *Natl. Acad. Sci*. U. S. A.

89(22), 10915–10919.

25. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., et al., (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res*. 49(D1), 412–419.
26. Stuart, A., Ord, K. (2010). Kendall's Advanced Theory of Statistics, Distribution Theory vol. 1. *John Wiley & Sons*.
27. Smith, T.F., Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol*. 147(1), 195–197.
28. Altschul, S.F., Madden, T.L., Sch¨affer, A.A., Zhang, J., Zhang, Z., et al., (1997). Gapped blast and psi-blast: a new generation of pro- tein database search programs. *Nucleic acids research* 25(17), 3389–3402.
29. Zhang, C., Shine, M., Pyle, A.M., Zhang, Y. (2022). Us-align: Universal struc- ture alignments of proteins, nucleic acids, and macromolecular complexes. *Bio Rxiv*.
30. Kunzmann, P., Hamacher, K. (2018). Biotite: a unifying open source compu- tational biology framework in Python. *BMC Bioinformatics* 19(1), 346.
31. Xu, J., Zhang, Y. (2010). How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics 26(7), 889–895.
32. Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: pat- tern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12), 2577–2637.
33. Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., et al., (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 43(Database issue), 364–368.