Received: 12 August 2022 | Revised: 30 September 2022 | Accepted: 3 October 2022

DOI: 10.1002/pro.4465

THE PROTEIN SOCIETY WILEY

DomainMapper: Accurate domain structure annotation including those with non-contiguous topologies

Edgar Manriquez-Sandoval[1] | **Stephen D. Fried**[1,2]

[1]T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA

[2]Department of Chemistry, Johns Hopkins University, Baltimore, MD, USA

Correspondence
Edgar Manriquez-Sandoval and Stephen D. Fried, T. C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA.
Email: emanriq1@jhu.edu, sdfried@jhu.edu

Funding information
National Institute of General Medical Science, Grant/Award Number: T32GM135131; National Science Foundation, Grant/Award Number: MCB-2045844

Abstract

Automated domain annotation is an important tool for structural informatics. These pipelines typically involve searching query sequences against hidden Markov model (HMM) profiles, yielding matches to profiles for various domains. However, domain annotation can be ambiguous or inaccurate when proteins contain domains with non-contiguous residue ranges, and especially when insertional domains are hosted within them. Here, we present DomainMapper, an algorithm that accurately assigns a unique domain structure annotation to a query sequence, including those with complex topologies. We validate our domain assignments using the AlphaFold database and confirm that non-contiguity is pervasive (10.74% of all domains in yeast and 4.52% in human). Using this resource, we find that certain folds have strong propensities to be non-contiguous or insertional across the Tree of Life. DomainMapper is freely available and can be ran as a single command-line function.

KEYWORDS

computational tools, domain prediction, domain topology, hidden Markov models

1 | INTRODUCTION

Proteins that perform complex functions are typically comprised of multiple semi-independently folding structural units called domains. These domains can be grouped together into categories that form the "periodic table" of the protein universe, elements that can be merged and combined in myriads of fashions. Automated domain detection has been critical for functional and structural annotation of proteins and plays important roles in protein structure prediction,[1–3] inference of protein function,[4–6] protein engineering, and rational truncation for protein expression, stabilization, and structural biology.

When using a large database for comprehensive domain annotation (such as SCOP2,[7,8] CATH-Gene3D,[9,10] Pfam,[11] CDD,[12] and HAMAP[13]), searches will typically yield many matches per protein, of which the majority will be redundant because a given residue range might match to many similar domains with various levels of confidence. Parsers play an important role in reducing this complexity by identifying a "minimal" set of domains such that a given residue range is at most assigned to one domain or two terminally overlapping domains; where conflicts arise, the "best" match is selected. Though it is simple to code a basic script for this task, conventional parsing strategies will exclude or mis-annotate domains with non-contiguous (NC) residue ranges, especially when there are insertional (IS) domains residing within them. We initially became intrigued with NC domains because recent studies found that *E. coli* proteins with NC domain topologies are generally less refoldable and rely more heavily on chaperones to assist their assembly.[14,15] These findings are consistent with the general understanding that long-range contacts in proteins are more challenging and slower to form.[16]

With the recent upgrade to v2.0,[17] every protein record in Uniprot possesses a Family & Domains section which draws data from InterPro,[18] an EMBL-EBI database that compiles 13 protein signature databases including CATH-

Gene3D, SCOP, and Pfam, among others. Whilst this integration of knowledge is powerful, the semi-overlapping nature of these annotation systems makes it challenging to assign a unique domain architecture to a protein. Moreover, Uniprot does not designate which domains are non-contiguous, insertional, or circularly permuted (CP). The biophysical relevance of NC domain topology, together with the relative inaccessibility of this information in the protein databases of record, prompted us to develop an easy-to-use annotation tool for this purpose.

To develop such a tool, we opted to focus on the domain definitions from ECOD,[19,20] though technically the program is compatible with domain definitions from any HMM-based protein signature database. ECOD is a frequently updated database with a comprehensive set of domain definitions (termed F-groups) that emphasize distant evolutionary relationships to construct a deep hierarchy (architecture > X-group > H-group > T-group > F-group) and has been particularly useful for computational structure prediction[21] and ancestral protein analysis.[22] Our program, DomainMapper, reads the alignments of matched ECOD domains in a query sequence and "fits" the domains (or pieces of a domain) together to generate a unique high-confidence domain map. Importantly, DomainMapper accurately assigns residue ranges to domains with non-trivial topologies, specifically NC and circularly permuted domains, as well as IS domains. We demonstrate the accuracy and generality of the algorithm by comparing the assigned domain ranges against predicted structures in the AlphaFold database using a few structural metrics. Using DomainMapper, we document the domain structures for the proteomes of eight model organisms from all domains of life, including humans. These annotations enable us to see clear trends of particular folds being significantly more amenable to being NC or IS over others. We expect that the accurate domain maps described here, and the ease-of-use of the program, will serve as a resource to the protein science community.

2 | DEVELOPMENT AND IMPLEMENTATION

2.1 | The DomainMapper algorithm

DomainMapper is supplied with a 'hmmscan -o' output file (Figure 1 Box 1),[23,24] and generates a unique maximum-confidence domain map for each query protein. In the output, each domain is assigned a residue range (that is possibly NC) and—if the ECOD domain database is used—is described by a set of labels in the ECOD structural hierarchy; specifically, an Architecture (e.g., a/b three-layer sandwich), an X-group (similar to a fold, for example, P-loop NTP hydrolase), a T-group (similar to a superfamily), and a F-group (the most specific classification, tied to a particular function). Proteins can be assigned multiple domains, which themselves can be assigned different topological annotations. Most domains are contiguous and are not given any labels, whilst others can be labeled NC, IS, circular permutant (CP), or a combination thereof.

To determine this assignment, the program sifts through each domain (represented as a Hidden Markov Model (HMM)) detected in a query sequence (called "hits"), which can match either once or multiple times, generating high-scoring pairs (HSPs). The simpler scenario occurs if a hit matches once (one HSP, Figure 1). After passing a confidence filter (conditional E-value $<10^{-5}$, an adjustable parameter), the match is appended to a list of potential domains to be considered during final reconciliation. Conditional E-values were used to minimize the effect of the search input's size. But first, the alignment between the HMM and the query is analyzed and for any gaps in the HMM longer than the intra-gap parameter (default is 30, the choice of which is discussed in Section 2.2), the corresponding residue indices in the query sequence are "carved out," giving rise to a discontinuous residue range for this potential domain and the opportunity for another domain to reside within these freed-up residue indices (Figure 1 Box 2).

The scenario is more complicated if there are multiple residue ranges in the query that match the model (i.e., multiple HSPs; Figure 1). In essence, this gives rise to three possibilities—for each pair of HSPs, they could either: (i) be overlapping (in which case the one with lower confidence should be eliminated); (ii) be separate and correspond to multiple copies of the same domain (e.g., as in a repeat-protein); or (iii) be separate but correspond to different portions of a single NC domain, in which case they should be merged together. Foremost, each HSP must still pass a confidence threshold (conditional E-value $<10^{-5}$) and all gaps (greater than intra-gap $= 30$, as above) in each alignment are carved out where found, as before. To ascertain which of the three possibilities is most appropriate, firstly, the optimal subset of non-overlapping HSPs within the query sequence is identified. If residue ranges in the query of the two HSPs overlap by more than the overlap parameter (default is 40), the one with the higher E-value (lower confidence) is eliminated, and the one with the lower E-value is retained. To ensure that the optimal set of HSPs is obtained, we implemented a recursive algorithm. This approach avoids the potential scenario in which a region with a weaker E-value is "mistakenly" eliminated, when it could have ultimately been retained if regions with multiple overlaps are eliminated first (Figure 1 Box 3). Because some HSPs are rather short, an overlap is also

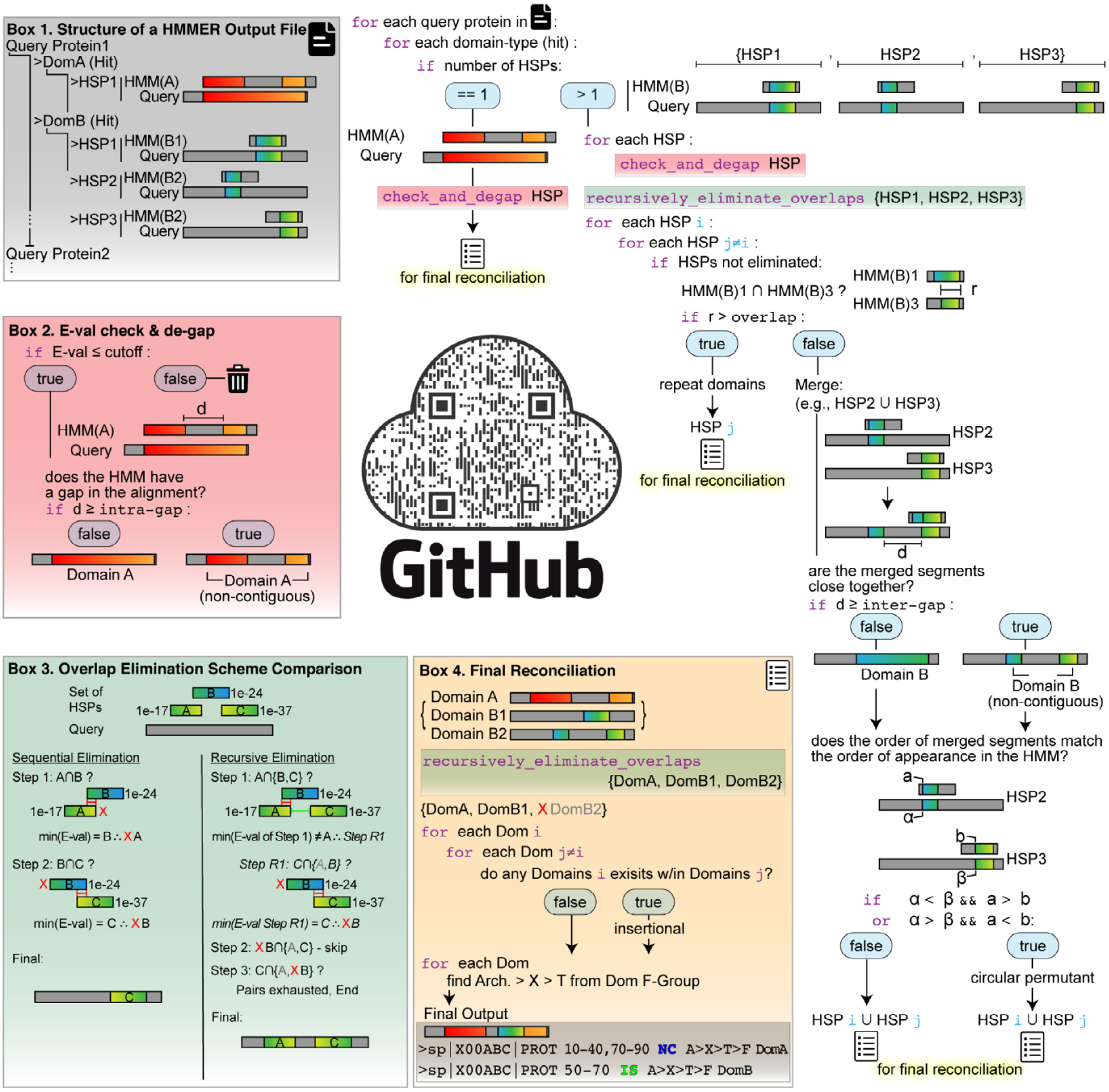

FIGURE 1 Flow chart describing DomainMapper's algorithm and functions for assigning protein domain structure. Box 1: Schematic representation of a HMMER output file, the input used by DomainMapper. Box 2: Algorithm for pre-processing HSPs based on *E*-value cutoffs and internal gaps in the aligned HMM sequences. Box 3: Comparison of two algorithms for eliminating overlapping domains. A sequential routine can result in incorrect eliminations, which DomainMapper mitigates by implementing a recursive algorithm. Box 4: Algorithm for final reconciliation in which protein structures are assessed based on combining results from each domain-type (hit)

called if it consists of 70% (or more) of the residues of either query (called the frac_overlap parameter, this too is adjustable).

Once the optimal set of non-overlapping HSPs is identified, we next determine which of these represent parts of a single (NC) domain or multiple copies of the same domain-type by iteratively analyzing HSPs in a pairwise manner. If the residue ranges in the HMMs of the two HSPs overlap (i.e., the same portion of the model matches two residue ranges of the query), then these are retained as two separate potential domains. If not (i.e., the two residue ranges in the query correspond to different portions of the model), then the two are merged, generating a NC domain. During mergers, *E*-values are combined using Fisher's method (as *E*-values and *p*-values are virtually identical for values less than .01). Most of the time, these two matching regions will be quite far from each other in sequence—hence, why they were identified as separate HSPs. In some cases, though, they are close together, and if they are closer than the inter-gap parameter (default is 30), the domain is "given" the residues in-between, reverting it to a contiguous domain. Essentially, the inter-gap parameter prevents a domain from being called NC simply because it was identified in two separate matches. Finally, for situations in which separate residue ranges were merged together, we ask if the order of these segments in the query matches the order of appearance in the HMM. If they do not, the domain is labeled as a

CP. After all mergers have been tested (and executed, if applicable), the matched regions are promoted to the set of potential domains, which accumulates mappings from all the different domain types (i.e., hits).

During final reconciliation (Figure 1 Box 4), all the potential domains that were discovered from all the different domain types "compete" for space in the query's residue space using the overlap parameter. In a pairwise manner, two potential domains are compared with one another. If the two domains overlap (more residues in common than overlap), the one with the higher *E*-value is eliminated; otherwise, both are retained. As discussed above, these comparisons are conducted using the same recursive algorithm as was applied to hits with multiple HSPs (though without mergers). After this process of elimination, a final highest-confidence domain structure annotation is produced, and each domain is labeled NC, if it possesses a discontinuity in its set of residue indices or IS, if its residue indices reside within another domain's.

2.2 | Assigning default parameters

The number and kind of NC and IS domains depend on the choice of the three parameters of DomainMapper: intra-gap, inter-gap, and overlap. Thus, we performed a sensitivity analysis on all three parameters in steps of five residues from 5–100 (for a total of $20^3 = 8{,}000$ parameter-sets) on the yeast proteome (6,079 proteins, isoforms included), and the results are given in Figure S1. The number of NC domains (N_{NC}) was most sensitive to intra-gap, very insensitive to inter-gap, and had a subtle dependency on overlap which depended on the value of intra-gap (Figure S1). The steep dependency of N_{NC} on intra-gap at low values makes sense in that when this parameter is too low, NCs will occur for every loop or small insertion in a query sequence's domain that was not present in the model. If the parameter is too high, gaps are "paved over" that could potentially host a domain. We chose 30 as the default value for intra-gap because: (i) this is close to the size of the smallest domains; (ii) this is at the inflection of N_{NC} (intra-gap) where exclusions switch regimes from the frequent loop-insertions to the less frequent domain-insertions.

N_{NC} shows very little dependency on inter-gap for the simple reason that the scenario in which it is invoked (stitching together two portions of a domain that matched as separate HSPs) is rare. Hence, applying the principal of parsimony, we assigned it the same value as intra-gap.

With the gap parameters established, we find that N_{NC} is only slightly sensitive to overlap (Figure S1), and so to choose its value, we instead looked at how it affects the number of IS domains (N_{IS}) and CP domains (N_{CP}) (Figure S1, respectively). Sensibly, it is easier to identify IS domains as overlap increases because this facilitates "fitting" inserts into spots where greater feathering occurs between the aligned regions' edges. We found that N_{IS} increases substantially (from 92 to 192) as overlap increases from 5 residues to 40, but then plateaus and slowly decreases. This decrease occurs because we make it easier to merge NC domains whose segments were detected as separate HSPs, thereby "paving over" regions that could host an insert. Consequently, setting overlap to 40 optimizes the number of inserts (Figure S1). As further evidence that this choice was sensible, it also minimizes $N_{\mathrm{NC}}/N_{\mathrm{IS}}$ whilst keeping N_{NC} high (Figure S1), in other words, it allows for detection of the most NC domains subject to the constraint that as many of these NC domains as possible are hosting IS domains, and not loops or other small insertions.

2.3 | Using the code

DomainMapper is freely available at https://github.com/FriedLabJHU/DomainMapper as a command-line tool written entirely in Python. Installation can be completed using pip and requires Python version 3.5 or newer and BioPython as dependencies.[25] DomainMapper can be executed within the terminal as dommap and should be provided a HMMER3 output file (generated with 'hmmscan -o') and the desired output path as inputs. Examples on how to use DomainMapper and instructions on using HMMER3 are provided in the README documentation on GitHub. Domain annotation of 100,100 human proteins (isoforms included), available on Uniprot, was completed in seven minutes on a single thread on an Intel i7-12700 processor with 20 cores and 32 GB of system memory.

3 | RESULTS AND DISCUSSION

3.1 | Testing the structural accuracy of NC domains

To check that the residue ranges which are assigned to NC domains by DomainMapper are accurate, we sought to evaluate these domains through high-throughput structural analyses (Figure 2).

For example, in the structure of human glutamate receptor GluR3 (Uniprot: P42263, Figure 2a), we find one periplasmic binding domain (#1, cyan and teal) becomes interrupted with a second periplasmic binding domain (#2, red). To illustrate the accuracy of the residue ranges assigned to these domains, we have painted them onto the AlphaFold structural prediction for GluR3 (Figure 2a). From visual inspection, one can see the centers of mass (CoMs, yellow spheres) of the two segments assigned to

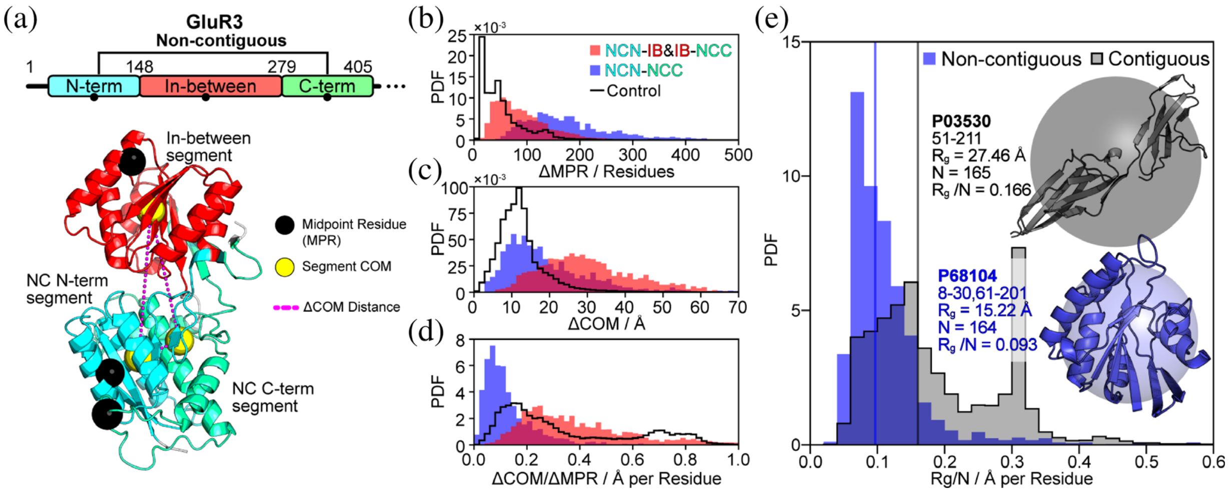

FIGURE 2 Accurate residue range calling of non-contiguous (NC) domains. (a) The N-terminal portion of human glutamate receptor 3 (GluR3, Uniprot: P42263) with a NC domain (N-terminal and C-terminal segments in cyan and teal, respectively) and an IS domain. The segment in between the two NC segments is shown in red. Black points mark the midpoint residue (MPR) for each segment. *Below*: The AlphaFold structural prediction of the GluR3 portion is shown above. Regions in cyan, red, and green correspond to the residue ranges of the three indicated segments. Each segment contains two spheres that represent the 3-D coordinates of the MPR (black) and center of mass (CoM, yellow). The magenta-dashed lines represent the distance between CoMs. (b–d) Distribution of the distances between segments (b) MPRs and (c) CoMs of all NC domains in *H. sapiens* ($N = 7{,}013$). In blue, distances between N-terminal and C-terminal NC segments (NCN, NCC respectively); in red, distances between the in-between (IB) segment and the two flanking NC segments. The distributions shown as black lines represent distances calculated from a set of controls in which all contiguous domains in *H. sapiens* ($N = 144{,}858$) are split into two halves, and the two halves are treated like the NCN and NCC segments of an NC domain. (d) As in panel C, except CoM-CoM distances are normalized by the two segments' MPR-to-MPR distance. (e) Distribution of the radius of gyration normalized to length (R_g/N) for domains in the *H. sapiens* proteome, either NC (blue) or contiguous (black) using domain ranges predicted by DomainMapper and structural models from AlphaFold. Lines represent medians (0.096 and 0.160 Å/residue, respectively). *Inset*: structures show a NC and contiguous domain selected from the *H. sapiens* proteome with R_g/N values representative of these two topology-types.

the NC domain are much closer in space to each other than either of them are to the center of mass of the "in-between (IB) segment," even though both NC segments are closer to the in-between segment in sequence space (which can be quantified as the difference between the "midpoint residue" [MPR, black spheres] of each region). For this analysis, we calculate the IB segment as all residues enveloped by the NC segments.

Leveraging the AlphaFold database of protein structure predictions, we can assess the generality of our approach by surveying 7,013 NC domains within the *H. sapiens* proteome (Figure 2b–d) and compare the prediction for the domain boundaries against the predicted structure. For each NC domain, we located the center of mass (CoM) and midpoint residue (MPR) for the segments that make up the NC domain as well as those for the regions that lie in-between (IB) using the residue ranges predicted by DomainMapper. Unsurprisingly, the MPR-to-MPR distance for NC-IB pairs (red) is generally closer to each other than the MPR-to-MPR distance for NC-NC pairs (blue), with median distances of 80 residues and 163 residues, respectively (Figure 2b). On the other hand, if we compare CoM-to-CoM distances (Figure 2c), it is apparent that NC-NC CoM-to-CoM distances are much shorter than NC-IB CoM-to-CoM distances, with the median (± std. dev.) distance of 16.2 ± 13.1 Å and 28.5 ± 13.3 Å, respectively. To further test DomainMapper's NC domain residue range annotations, we compared the NC-NC CoM-to-CoM distance distribution (blue) to a control set (black line), in which normal contiguous domains were split in half, and CoMs were assigned to the two resulting segments. Satisfyingly, we found that CoM-to-CoM distances were similar among actual NC domains (median, 16.2 ± 13.1 Å) compared to these controls (median, 11.5 ± 7.6 Å). When we normalized Euclidean CoM-to-CoM distances with the 1-D MPR-to-MPR distances (Figure 2d)—a metric that adjusts for the inherent tendency of segments that are close in sequence to be close in space—we find that the distribution of these normalized distances for NC-NC (blue) pairings is significantly shorter than for the control (black) pairings (Figure 2d, medians 0.100 Å/residue and 0.249 Å/residue respectively, $p = 0$ by the Mann–Whitney rank-sum test). This shows that the residue ranges assigned to NC domains create compact globular structures even though they can be quite far apart in sequence.

To further demonstrate the accuracy of DomainMapper's NC domain residue range annotations, we calculated the radii of gyration (normalized to domain length, R_g/N) of all NC domains in *H. sapiens* and compared them to the R_g/N of contiguous domains. Accurate residue ranges for NC domains, we hypothesized, would generate a distribution of R_g/N comparable to that of contiguous domains. Strikingly, we found that R_g/N is substantially smaller on average for NC domains (median 0.096 Å/residues) than it is for contiguous domains (median 0.160 Å/residues; $p = 0$ by the Mann–Whitney rank-sum test; Figure 2e). The inset of the figure shows an example NC and contiguous domains with representative R_g/N values equal to each group's median. Hence, we conclude that DomainMapper accurately assigns residue ranges to NC domains and report the finding that NC domains are generally more structurally compact than average. Moreover, all these findings held when the same structural analyses were applied to the yeast proteome as well (Figure S2).

3.2 | Assessment of domain structures across several model proteomes

We find that NC domains (and IS domains) are universal (Figure 3). In humans, 4.52% of all domains have a NC residue range ($N = 7{,}013$). In yeast, this fraction rises to 10.74% ($N = 785$). IS domains are typically 1.5–4 times less frequent (1.72% of all domains in humans ($N = 2{,}661$), 2.63% in yeast ($N = 192$)). This is because NC domains are sometimes interrupted by linker regions rather than by other folded domains. Interestingly, in the prokaryotes, we surveyed (especially *S. aureus*), NC domains are more likely to host IS domains than in most eukaryotes. This observation possibly follows from the fact that eukaryotic proteomes are typically more disordered in general, and so the same is true for regions within domains as for regions outside domains.

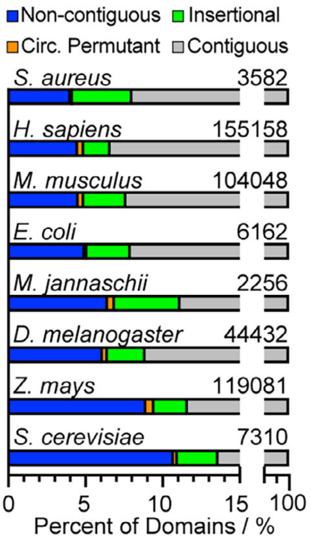

FIGURE 3 Proteome-wide domain annotations of eight model organisms. Bar charts showing the frequency of topological labels assigned to individual domains by DomainMapper: non-contiguous (NC, blue); circular permutant (CP, orange); insertional (IS, green); and contiguous (CON, gray). *S. aureus*, $N = 3{,}582$ (NC: 4.05%, CP: 0.14%, IS: 3.85%, CON: 91.96%); *H. sapiens*, $N = 155{,}158$ (NC: 4.52%, CP: 0.40%, IS: 1.72%, CON: 93.36%); *M. musculus*, $N = 104{,}048$ (NC: 4.57%, CP: 0.35%, IS: 2.74%, CON: 92.35%); *E. coli*, $N = 6{,}162$ (NC: 4.98%, CP: 0.13%, IS: 2.82%, CON: 92.06%); *M. jannaschii*, $N = 2{,}256$ (NC: 6.47%, CP: 0.44%, IS: 4.26%, CON: 88.83%); *D. melanogaster*, $N = 44{,}432$ (NC: 6.13%, CP: 0.33%, IS: 2.44%, CON: 91.10%); *Z. mays*, $N = 119{,}081$ (NC: 8.95%, CP: 0.53%, IS: 2.18%, CON: 88.34%); *S. cerevisiae*, $N = 7{,}310$ (NC: 10.74%, CP: 0.26%, IS: 2.63%, CON: 86.37%)

3.3 | DomainMapper captures cases with complex domain topology

The most common type of protein with a discontinuous domain topology is that which possesses one NC domain with two segments. However, DomainMapper can also describe proteins with complex domain structures. In Figure 4, we provide two examples. In the automated domain assignment generated by DomainMapper for the β' subunit of RNA polymerase (RpoC, Uniprot: Q8RQE8 from *Thermus thermophilus*, PDB: 6KQG,[26] Figure 4a), RpoC has an unusually complex domain structure, consisting of 12 distinct structural domains. Three of these domains are NC (domain 1 in red, domain 6 in lime and domain 11 in purple) and six are IS (domains 2, 3, 4, 5, 7, and 12). The algorithm properly recapitulates the complex domain topology of this protein, assembling together the pairs of segments that form the N-terminal domain (1), the central helical domain (11), and the double-psi cradle loop barrel (6).

DomainMapper also describes NC domains with higher numbers of disconnected segments. *E. coli*'s leucyl-tRNA synthetase (LeuS, Uniprot: P07813, PDB: 4AQ7,[27] Figure 4b) contains an unusual class I aminoacyl-tRNA synthetase domain with five separate segments, a situation that occurs only once in *E. coli*. Whilst NC domains with three segments are not the norm (Figure 4c; 27 in total in *E. coli*, 129 in *S. cerevisiae*), they are also not extremely rare. On the other hand, NC domains with more than five segments do not occur in these two organisms.

3.4 | Non-contiguity is systematically enriched in different folds

In the ECOD formalism, all domains are assigned a position in a hierarchy based on evolutionary relationships,

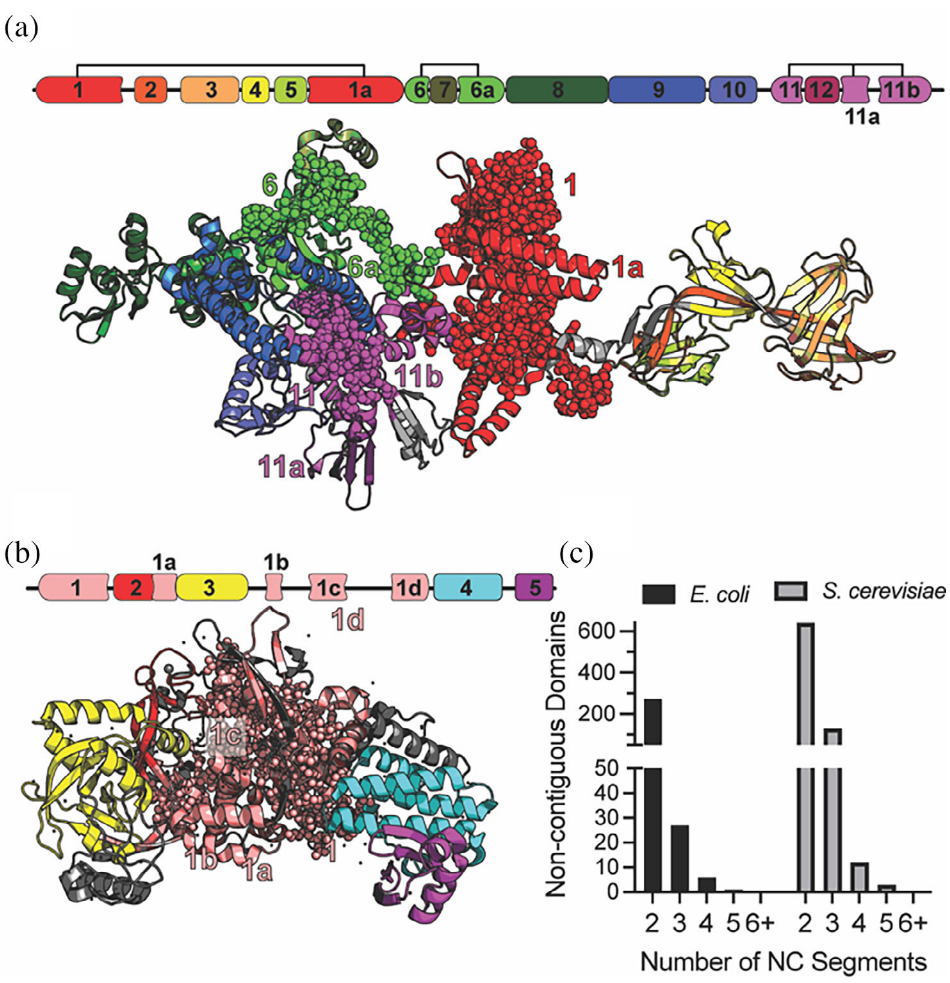

FIGURE 4 Accurate annotation of complex domain topologies. (a) DomainMap of *Thermus thermophilus* β' subunit of RNA polymerase (RpoC, Uniprot: Q8RQE8) with 12 domains. Segments of non-contiguous (NC) domains are labeled with letters. One segment of each NC domain is represented as spheres to distinguish it from the other segments. (b) DomainMap of *E. coli* leucyl-tRNA synthetase (LeuS, Uniprot: P07813) with five domains and one NC domain with five segments. (c) Bar chart showing the number of NC domains in the *E. coli* and *S. cerevisiae* proteomes with the indicated number of segments.

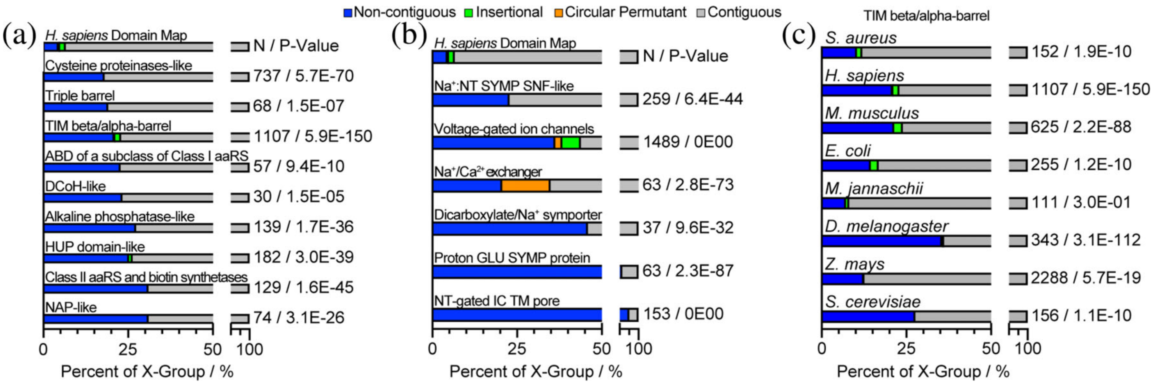

FIGURE 5 Non-contiguity is enriched in certain fold-types. (a) Bar charts showing the frequency of NC (blue), circular permutant (CP) (orange), insertional (IS) (green), and contiguous (gray) domains across the human proteome for several X-groups that are enriched for non-contiguity relative to other human protein domains. Also given are the number of domains in that X-group identified in total (N), and the *p*-value (by chi-square test) that the over-/under-representation of these labels within the given X-group, relative to the overall human proteome, are not due to chance. (ABD, anticodon-binding domain; NAP, nucleosome assembly protein). (b) Analogous to panel A, but for several specialized X-groups related to channels and transporters that are very enriched for non-contiguity. (NT, neurotransmitter; SYMP, symport). (c) Bar chart showing the frequency of non-contiguity of the TIM barrel X-group in eight species. In all species, TIM barrels are disproportionately NC except in *M. jannaschii*

termed X-groups, T-groups, and F-groups. These classifications correspond loosely to folds (X-groups), superfamilies (T-groups), and families (F-groups) from the SCOP classification system. Using this formalism, we counted the number of times domains in each X-group had a non-standard topology (e.g., NC, IS, or circularly permuted), and found that these properties are unevenly spread across protein domain space (Figure 5).

Using the *H. sapiens* proteome as a model, we found amongst the universal folds, TIM barrels, cysteine protease-like domains, and alkaline phosphatase-like folds are significantly more likely to be NC: specifically, 4.6-fold, 4-fold, and 6-fold more frequently than average (4.52%) (Figure 5a). These enrichments are highly statistically significant in relation to their expected frequencies based on the chi-square test (*p*-values of 6×10^{-150}, 6×10^{-70}, and 2×10^{-36}, respectively). From biophysical studies, TIM barrels and alkaline phosphatase are generally noted for their great stability.[28–31] It would seem that this stability could have served as an important foundation to make these folds more amenable to accommodating large insertions within their domain.

We also found an unusually strong preference for folds associated with aminoacyl-tRNA synthetases (aaRS) to be NC (Figure 5a). This corresponds to the X-groups for the class II aaRS core fold, HUP domains (which contain the core class I aaRS fold), and an anticodon binding domain associated with class I aaRS. For these categories, non-contiguity is enriched 6.9-fold, 5.6-fold, and 5-fold, respectively. The high propensity of non-contiguity in these enzymes is consistent with their general inability to reversibly refold,[14,15] and is consistent with the conjecture that they co-evolved with the emergence of protein synthesis by translation.[32]

Several X-groups associated with integral membrane transporters are quite enriched for non-contiguity (Figure 5b),[33] including neurotransmitter-gated ion-channel folds (17-fold), proton glutamate symport folds (13-fold), and voltage-gated ion channels (8-fold). These X-groups are highly attested in the human proteome for which recent gene duplication and many paralogs are available. Hence, these results suggest that NC topologies play an important role in enabling signal-recognizing domains to communicate allosterically with transmemb rane-regions during transport.

The trends that we have described here, though based on the human proteome, are generally consistent across the species for which we have mapped domains (Figure 5c). For instance, in all organisms that we surveyed (except for *M. jannaschii*), TIM barrels are more likely to have NC topologies. Likewise, this enrichment is also found for the aaRS-associated folds across the eight model proteomes analyzed. These observations suggest that the enrichment of non-contiguity in these particular folds is not species-specific, but rather reflect their inherent biophysical qualities.

3.5 | Certain folds are enriched as insertional domains

We also carried out an analysis to determine how the IS status is distributed across the human proteome (Figure 6). In general, IS domains are comparatively rarer (2–4%, cf. Figure 3), implying that it is generally more facile to create multi-domain proteins by appending (or prepending) domains to a protein in a tail-to-head manner than to interrupt a domain with another one. That said, certain X-groups appear more amenable to be inserted (Figure 6a), notably: cradle loop barrels, FKBP-like domains, and rubredoxin-like domains (for which the IS status is enriched 2.9-fold, 6-fold, and 10-fold,

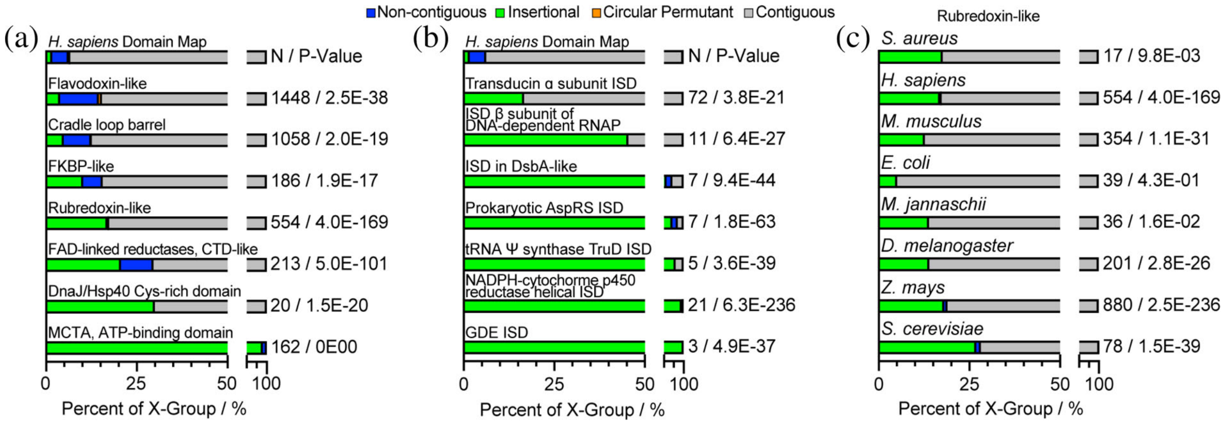

FIGURE 6 Certain fold-types are more likely to occur as inserts. (a) Bar charts showing the frequency of NC (blue), CP (orange), IS (green), and contiguous (gray) domains across the human proteome for several X-groups that are enriched to be IS relative to other human protein domains. Also given are the number of domains in that X-group identified in total (N), and the P-value (by chi-square test) that the over−/under-representation of these labels within the given X-group, relative to the overall human proteome, are not due to chance. (CTD, C-terminal domain; MCTA, metal cation-transporting ATPase). (b) Analogous to panel A, but for X-groups that are annotated as insertional domains (ISD). (GDE, glycogen debranching enzyme). (c) Bar chart showing the frequency of being IS of the Rubredoxin-like X-groups in eight species. All species demonstrate significant enrichment for Rubredoxin-like insertionality except for *E. coli*

respectively). These X-groups contain relatively small folds (≤70 residues) with N- and C-termini facing in similar directions, which would minimize disruption to the host NC domains.

Tellingly, DomainMapper found a handful of X-groups that are ultra-enriched to be inserted (with several over 50%), and satisfyingly all of these fold-types are annotated as insertional domains (ISD, Figure 6b), implying that DomainMapper recapitulates this feature. In these cases, it is likely that the domain in question arose originally as an insert within a host domain, and that the two-domain cassette served as a common ancestor for further appearances of the inserted domain.

We encountered several folds that are systematically deprived of being IS. Some of these are small all-α domains that are quite extended (i.e., where the N- and C- termini point away from each other), which would readily explain why they would be difficult to insert on geometrical grounds (notably, the EF-hand, histone-like, and repetitive α-hairpins). Finally, we find that most of the trends described are conserved across species; for instance, rubredoxin has a significant preference to be inserted across all proteomes we studied except for *E. coli* (Figure 6c).

3.6 | Comparison to CATH-Gene3D

In principle, the ideal way to test DomainMapper's performance at detecting NC domains would be a sensitivity-specificity (ROC) analysis. This is not possible, however, because there is no ground truth of what constitutes an NC domain; hence, an agreed-upon test set is not presently available. In looking for potential validation strategies, we found that discontinuous non-overlapping residue ranges for CATH-Gene3D domains could be obtained by querying the InterPro API, enabling a side-by-side comparison of two distinct available methods to identify protein domain maps and NC domains in particular. Applied to the 6,079 proteins in the yeast proteome, we found that CATH-Gene3D and DomainMapper yield similar levels of coverage (6,927 non-overlapping domains found by CATH-Gene3D; cf. 7,310 from DomainMapper). CATH-Gene3D's internal parser identified 907 NC domains (cf. 785 from DomainMapper), which we examined using the structural metrics described in Section 3.1 (Figure S3). On the whole, the distribution of normalized CoM-to-CoM distances for NC-NC pairings was similar between CATH-Gene3D's and DomainMapper's NC domain residue ranges (Figure S3A), and the distribution of normalized radii of gyration of NC domains was nearly identical (Figure S3B). This shows that CATH-Gene3D's internal parser is also able to accurately call discontinuous residue ranges, and generates NC domains that are structurally compact, like DomainMapper.

On the other hand, we found that among the 907 NC domains identified by CATH-Gene3D, only 353 of these were also found with DomainMapper (and the remaining 554 were unique to CATH-Gene3D only). Upon examining the NC domains unique to CATH-Gene3D, we saw that many of these NC domains had very short in-between regions (Figure S4A) consisting of loops or small insertions—analogous to DomainMapper's behavior for smaller values of intra-gap (see Figure S1). When the intra-gap and inter-gap parameters of DomainMapper are set to 10 residues, we find that DomainMapper identifies the majority (60%) of CATH-Gene3D's NC domains, and moreover identifies twice as many NC domains overall (Figure S4B). During our parameter optimization studies, we assigned intra-gap a default value of 30 to engender a definition of a NC domain that focuses on cases in which other domains (or domain-sized linkers) are inserted. NC domains uniquely identified by DomainMapper appeared as bona fide cases of non-contiguity by manual validation (Figure S4A). Our assessment is that CATH-Gene3D's internal parser is quite good; nevertheless, the default parameters used in DomainMapper assign NC status in a manner closer to “protein structure intuition.”

We note that InterPro's API did not provide any discontinuous residue ranges for Pfam domains, precluding a comparison.

Finally, we performed a short comparative study on a set of other previously developed domain annotation tools, as compiled by Xue and co-workers.[34] Our attempt to test 19 previously developed sequence-based algorithms yielded 15 that are no longer available or accessible, and four (ThreaDom, MetaCLADE, DeepDom, FuPred) that were. Three of these (ThreaDom, FuPred, and DeepDom) tools predict possible domain ranges but do not annotate them. The other one (MetaCLADE) requires a pre-curated set of Pfam domains to be considered. Hence, DomainMapper possesses a constellation of factors that in our view make it a more useful tool, and by being fully open-source and available on Github, its availability should be assured into the future.

4 | CONCLUSION AND FUTURE DIRECTION

Protein domain annotation is an important part of structural biology and consists of two distinct tasks: (1) detecting domains by aligning signatures (typically HMM profiles) to query sequences; and (2) parsing the results to obtain a highest-confidence non-overlapping domain map. There are several different (and complementary) HMM profile databases available for the first task (e.g., SCOP2, Pfam, CATH, ECOD, etc.) with distinct advantages and disadvantages depending on the user's question; hence, we do not think uniformity for that task is needed or desired. In this respect, the compilation of multiple protein signature databases in Uniprot and InterPro is excellent resources for the protein science community. On the other hand, we believe there is a need for a general, flexible, and scalable algorithm to parse identified domains into unique high-confidence domain maps with allowances for complex domain topologies—a need that we have endeavored to address with DomainMapper.

With this tool available, ECOD's online database (http://prodata.swmed.edu/ecod/) could be rendered more user-friendly by displaying domain maps rather that raw HMMER output upon performing sequence search in a browser. However, even though we focused on the ECOD set of domain definitions for this study, DomainMapper can accept inputs from any HMM-based protein signature database. As a longer-term goal, we recommend that the Family & Domains section on each protein's Uniprot page be simplified, so that instead of presenting a list of matching domains (the outcome of task 1), domain maps of the protein for each system could be presented instead. It is possible that the reason this has not already been done is that different protein signature databases developed custom parsers, which have not been made publicly available. We seek here to lift that roadblock.

As with all HMM-based domain prediction tools, DomainMapper's precise residue ranges for each domain do not always agree with structural models, an imprecision that occurs because the alignments of query sequences to HMMs tend to be weaker at the beginnings and ends. Protein domain annotation could clearly benefit from the meteoric advances in computational structure prediction, in which residue ranges could be amended by placing the initial annotation into a structural model, interrogating the immediate upstream and downstream regions, and adjusting the annotation accordingly. Such a development would most likely improve DomainMapper further.

Using DomainMapper, we were able to reliably annotate complex topologies such as NC domains, IS domains, and CP domains across a small number of model proteomes. Our analysis shows that certain fold-types are enriched with these topological statuses, observations that correlate with their biophysical or structural properties. It is possible that non-contiguity also has *functional* consequences, enabling the activities of two domains to be more closely coupled than would be possible were the domains tethered to one another on only one end. For instance, aminoacyl-tRNA synthetases must closely couple recognition of the anticodon loop of tRNA with aminoacylation on the distal tip of the acceptor arm, and ligand/voltage-gated transporters must closely couple conformational changes in the ligand/voltage-recognizing domain with channel opening/closing (cf. Figure 5). Hence, it is possible that NC domain topology is associated with certain more complex biochemical functions. We expect that that more patterns will emerge as DomainMapper is applied to analyze domains more deeply across the Tree of Life and in the service of other bioinformatic questions.

5 | METHODS

5.1 | Resources used in this study

Reagent or resource	Source	Identifier
Deposited data		
Domain mappings for all proteomes	This paper	10.5281/zenodo.6980883
All proteomes with canonical and isoforms proteins	This paper	10.5281/zenodo.6980786
Software and algorithms		
HMMER3 v3.3.2	Ref. 23	http://hmmer.org
Python v3.9.12	Python software foundation	https://www.python.org
Pymol molecular graphics system	Schrodinger, LLC	https://pymol.org
Biopython v1.60	Biopython.org	https://biopython.org
Numpy v1.17.0	Numpy.org	https://numpy.org
SciPy v1.6.1	SciPy.org	https://scipy.org/
ECOD domain definitions	Ref. 19	http://prodata.swmed.edu/ecod/
DomainMapper 3.0.1	This paper	10.5281/zenodo.6981403 https://github.com/FriedLabJHU/DomainMapper
Domain mapper companion analyses (high-throughput structural analyses)	This paper	https://github.com/FriedLabJHU/DomainMapper-Companion-Analyses
All proteome AlphaFold 3D structures	Ref. 2	https://alphafold.ebi.ac.uk/download

5.2 | Data and code availability

The data reported in this study were generated with DomainMapper v3.0.1, which have been deposited and as of the date of publication is publicly available here: source code (10.5281/zenodo.6981403) and data (10.5281/zenodo.6980883). Future releases of the DomainMapper software will be available at https://github.com/FriedLabJHU/DomainMapper. Domain maps created with DomainMapper v3.0.1, hidden Markov model domain alignments created with HMMER3 v3.3.2, and proteome sequence FASTA files accessed from Uniprot on August 9, 2021, November 1, 2021, and July 8, 2022, have been deposited at (10.5281/zenodo.6980786) and are publicly available as of the date of publication. DOIs are listed in the resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

5.3 | Computational methods

5.3.1 | Domain alignments utilizing ECOD domain definitions in HMMER3

The ECOD formalism is not a standard hidden Markov model (HMM) profile available on the HMMER3 webserver and utilization thus requires preparation of the ECOD family HMM profile on a local machine. The prebuilt HMM profiles are available for download from http://prodata.swmed.edu/ecod/ecodf.hmm.tar.gz, which are regularly updated on a 2–3-month basis. Processing of the ECOD HMM profile was performed in HMMER3 v3.3.2 following a process available in the software documentation summarized elsewhere.[23,24] The extracted ECOD HMM profile download is passed to the hmmpress HMMER3 function, which recompresses the profiles into a proprietary binary file and index file that can be utilized to create domain alignments. Domain alignments were performed using the hmmscan HMMER3 function with the ECOD binary files and proteome FASTA files as inputs. Since DomainMapper requires the full output from hmmscan (as opposed to truncated outputs generated by the -tblout and -domtblout flags), the -o flag was set to true and provided the desired output file name. In the output generated by hmmscan, every protein is specified by a query name and Uniprot accession label followed by a summary of the domains detected in the query protein sequence along with their residue ranges. For each hit, a series of residue ranges in the query protein that match it (high-scoring pair, HSP) are provided. For each HSP, an alignment of the query sequence to the HMM model is provided along with several goodness-of-fit metrics (such as score and conditional *E*-value) and residue ranges, both with respect to the query protein and the HMM. Proteomes were accessed from Uniprot on August 9, 2021, and November 1, 2021, for the eight model organisms reported in this study and aligned in the same way. HMM scanning the 100,100 proteins (isoforms included) of the *Homo sapiens* proteome (Uniprot: UP000005640) took 18 hours on an Intel i7-12700 processor with 20 cores and 32 GB of system memory in parallel.

5.3.2 | The DomainMapper program

Using default settings, DomainMapper's command-line tool, dommap, only requires a HMMER3 output file and preferred output directory to produce domain annotations. The three parameters mentioned herein, intra-gap, inter-gap, and overlap, are adjustable by the user via the use of the --intra_gap, --inter_gap, and --overlap flags, respectively. Additionally, the *E*-value cutoff for domains to be considered for mapping can be adjusted by the --eval_cutoff flag. Upon initial installation, the latest ECOD domains definitions, available at http://prodata.swmed.edu/ecod/ecod.latest.domains.txt, are parsed and formatted to facilitate the bottom-up search of the domain structural hierarchies (F-group, T-group, H-group, X-group, Architecture). The ECOD domain definitions can be manually updated with the --update flag; however, dommap will automatically update them every two months. To utilize domain definitions and structural hierarchies of other classifiers (e.g., CATH, SCOP, Pfam, etc.), the --dom_def flag can be provided a text file, formatted similar to ecod.latest.domains.txt. Mapping of the 155,158 domains of the *Homo sapiens* proteome, with the default dommap settings, took 7 min on a single thread on an Intel i7-12700 processor with 20 cores and 32 GB of system memory.

5.3.3 | High-throughput structural analyses

For all NC domains in the Yeast and Human proteomes, the DomainMapper output was parsed for NC domain residue ranges. The centers-of-mass and midpoint residues were located in both the N-terminal and C-terminal NC segments as well as the interior segment between the non-contiguous segments, the so-called in-between (IB) segment. The normalized center-of-mass distance was defined by the Euclidian distance between centers-of-mass of two segments divided by the absolute difference between the segments' midpoint residues. A control population was also analyzed by taking all contiguous domains and randomly splitting them into segments in 30–70% fractions and performing the same calculations as with the non-contiguous domains except without the in-between segments. Centers-of-mass were calculated with Biophython (Bio.PDBParser) for all segments from AlphaFold predicted structures where available. The Euclidian distance between centers-of-mass and the absolute difference between midpoint residues were calculated with NumPy and separated into populations of NC-NC pairs and NC-IB pairs. Radii of gyration were calculated with NumPy using the atomic coordinates from structures and atomic masses, except the calculations were conducted on whole domains (either contiguous or NC) rather than on separate segments. To perform comparisons to CATH-Gene3D, calculations were performed identically except using residue ranges (contiguous or otherwise) from CATH, as described below. The available AlphaFold structures were downloaded from the EBI bulk download distribution. These analyses were conducted and visualized using Jupyter notebooks and are available at https://github.com/FriedLabJHU/DomainMapper-Companion-Analyses.

5.3.4 | Comparison to CATH

For all accessions in the Yeast DomainMapper output, CATH-Gene3D domain residue ranges were obtained using the Python library requests v2.28.1. HTTPS requests were made to the InterPro API, https://www.ebi.ac.uk/interpro/wwwapi//entry/all/protein/reviewed/ followed by the Uniprot accession code, returning a JSON file containing domain metadata. Domain start and end ranges were parsed from the "fragments" field in the metadata. Domains containing multiple "fragments" entries were labeled NC; otherwise, they were labeled as contiguous domains. For domains where the NC assignment agreed in both DomainMapper and CATH-Gene3D, a count was maintained as belonging to both; otherwise, when the NC assignment was unique to either of the annotators, a separate count was tallied. These residue ranges were additionally used to perform the structural analyses shown in Figure S3. These analyses were conducted and visualized using Jupyter notebooks and are available at https://github.com/FriedLabJHU/DomainMapper-Companion-Analyses.

5.3.5 | Statistical analyses

All statistical analyses were performed with Python v3.9.12. The chi-square tests in Figures 5 and 6 were performed with the Python library SciPy v1.6.1 and significance was determined for p-values <.05. The total domain counts in Figure 3 are provided by DomainMapper in the output header, and the total X-group counts in Figures 5 and 6 were calculated with the Python library NumPy v1.17.0. The two-sided Mann–Whitney rank sum tests used to compare R_g/N distributions and ΔCoM/ΔMPR distributions were all performed in SciPy with the pre-built methods. When P-values are reported as zero, it is because they are too small to be represented by these Python libraries ($<10^{-300}$).

AUTHOR CONTRIBUTIONS

Edgar Manriquez-Sandoval: Formal analysis (lead); investigation (lead); software (lead); writing – original draft (supporting). **Stephen Fried:** Conceptualization (lead); funding acquisition (lead); project administration (lead); software (supporting); writing – original draft (lead).

ACKNOWLEDGMENTS

We would like to thank the NSF Division of Molecular and Cellular Biology (2045844) for support. EMS thanks the Program in Molecular Biophysics training grant (NIH T32GM135131). We also thank Liam Longo for advice and suggestions about how to use the ECOD database.

CONFLICTS OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Domain mappings for all proteomes: 10.5281/zenodo.6980883; All proteomes with canonical and isoforms proteins: 10.5281/zenodo.6980786; DomainMapper 3.0.1: 10.5281/zenodo.6981403; https://github.com/FriedLabJHU/DomainMapper; Domain Mapper Companion Analyses (High-throughput structural analyses): https://github.com/FriedLabJHU/DomainMapper-Companion-Analyses.

ORCID

Edgar Manriquez-Sandoval https://orcid.org/0000-0001-7284-1237
Stephen D. Fried https://orcid.org/0000-0003-2494-2193

REFERENCES

1. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science. 2003;300:1701–1703.
2. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.
3. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol. 2004;14:208–216.
4. Friedberg I. Automated protein function prediction—The genomic challenge. Brief Bioinform. 2006;7:225–242.

5. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007;812:995–1005.
6. Osadchy M, Kolodny R. Maps of protein structure space reveal a fundamental relationship between protein structure and function. Proc Natl Acad Sci. 2011;108:12301–12306.
7. Pandurangan AP, Stahlhacke J, Oates ME, Smithers B, Gough J. The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. Nucleic Acids Res. 2019;47: D490–D494.
8. Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020;48:D376–D382.
9. Sillitoe I, Dawson N, Lewis TE, et al. CATH: Expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 2019;47:D280–D284.
10. Sillitoe I, Bordin N, Dawson N, et al. CATH: Increased structural coverage of functional space. Nucleic Acids Res. 2021;49: D266–D273.
11. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–D432.
12. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res. 2020;48:D265–D268.
13. Pedruzzi I, Rivoire C, Auchincloss AH, et al. HAMAP in 2015: Updates to the protein family classification and annotation system. Nucleic Acids Res. 2015;43:D1064–D1070.
14. To P, Whitehead B, Tarbox HE, Fried SD. Nonrefoldability is pervasive across the *E. coli* proteome. J Am Chem Soc. 2021; 143:11435–11448.
15. To P, Xia Y, Devlin T, Fleming KG, Fried SD. A proteome-wide map of chaperone-assisted protein refolding in a cellular-like milieu. bioRxiv. 2022. https://doi.org/10.1101/2021.11.20.469408v2.
16. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol. 1998;277:985–994.
17. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acis Res. 2021;49:D480–D489.
18. Blum M, Chang H-Y, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–D354.
19. Cheng H, Schaeffer RD, Liao Y, et al. ECOD: An evolutionary classification of protein domains. PLoS Comput Biol. 2014;10: e1003926.
20. Schaeffer RD, Kinch LN, Liao Y, Grishin NV. Classification of proteins with shared motifs and internal repeats in the ECOD database. Protein Sci. 2016;25:1188–1203.
21. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics. 2018;34:3308–3315.
22. Longo LM, Jablonksa J, Vyas P, et al. On the emergence of P-loop NTPase and Rossmann enzymes from a Beta-alpha-Beta ancestral fragment. eLife. 2020;9:e64415l.
23. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
24. Eddy SR. Hidden Markov models. Curr Opin Struct Biol. 1996; 6:361–365.
25. Cock PJA, Antao T, Chang JT, et al. Biopython: Freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–1423.
26. Li L, Molodtsov V, Lin W, Ebright RH, Zhang Y. RNA extension drives a stepwise displacement of an initiation-factor structural module in initial transcription. Proc Natl Acad Sci USA. 2020;117:5801–5809.
27. Palencia A, Crepin T, Vu MT, Lincecum TL Jr, Martinis SA, Cusack S. Structural dynamics of the Aminoacylation and proofreading functional cycle of bacterial Leucyl-tRNA Synthetase. Nature Struct Mol Biol. 2012;9:677–684.
28. Dong G, Zeikus JG. Purification and characterization of alkaline phosphatase from Thermotoga neapolitana. Enzyme Microb Technol. 1997;21:335–340.
29. Romero-Romero S, Costas M, Rodríguez-Romero A, Fernández-Velasco D. Reversibility and two state behaviour in the thermal unfolding of oligomeric TIM barrel proteins. Phys Chem Chem Phys. 2015;17:20699–20714.
30. Sterner R, Höcker B. Catalytic versatility, stability, and evolution of the $(\beta\alpha)_8$-barrel enzyme fold. Chem Rev. 2005;105:4038–4055.
31. Zees AC, Pyrpassopoulos S, Vorgias CE. Insights into the role of the $(\alpha + \beta)$ insertion in the TIM-barrel catalytic domain, regarding the stability and the enzymatic activity of Chitinase a from Serratia marcescens. Biochim Biophys Acta—Proteins Proteom. 2009;1794:23–31.
32. Fried SD, Fujishima K, Makarov M, Cherepashuk I, Hlouchova K. Peptides before and during the nucleotide world: An origins story emphasizing cooperation between proteins and nucleic acids. J Roy Soc Interface. 2022;19:20210641.
33. Coyote-Maestas W, Nedrud D, Suma A, et al. Probing ion channel functional architecture and domain recombination compatibility by massively parallel domain insertion profiling. Nature Commun. 2021;121:1–16.
34. Wang Y, Zhang H, Zhong H, Xue Z. Protein domain identification methods and online resources. Comp Struct Biotech J. 2021;19:1145–1153.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Manriquez-Sandoval E, Fried SD. DomainMapper: Accurate domain structure annotation including those with non-contiguous topologies. Protein Science. 2022; 31(11):e4465. https://doi.org/10.1002/pro.4465