AlQuraishi *BMC Bioinformatics* (2019) 20:311
https://doi.org/10.1186/s12859-019-2932-0

BMC Bioinformatics

DATABASE Open Access

ProteinNet: a standardized data set for machine learning of protein structure

Mohammed AlQuraishi

Abstract

Background: Rapid progress in deep learning has spurred its application to bioinformatics problems including protein structure prediction and design. In classic machine learning problems like computer vision, progress has been driven by standardized data sets that facilitate fair assessment of new methods and lower the barrier to entry for non-domain experts. While data sets of protein sequence and structure exist, they lack certain components critical for machine learning, including high-quality multiple sequence alignments and insulated training/validation splits that account for deep but only weakly detectable homology across protein space.

Results: We created the ProteinNet series of data sets to provide a standardized mechanism for training and assessing data-driven models of protein sequence-structure relationships. ProteinNet integrates sequence, structure, and evolutionary information in programmatically accessible file formats tailored for machine learning frameworks. Multiple sequence alignments of all structurally characterized proteins were created using substantial high-performance computing resources. Standardized data splits were also generated to emulate the difficulty of past CASP (Critical Assessment of protein Structure Prediction) experiments by resetting protein sequence and structure space to the historical states that preceded six prior CASPs. Utilizing sensitive evolution-based distance metrics to segregate distantly related proteins, we have additionally created validation sets distinct from the official CASP sets that faithfully mimic their difficulty.

Conclusion: ProteinNet represents a comprehensive and accessible resource for training and assessing machine-learned models of protein structure.

Keywords: Proteins, Protein structure, Machine learning, CASP, Protein sequence, Co-evolution, PSSM, Protein structure prediction, Database, Deep learning

Background

Deep learning has revolutionized many areas of computer science including computer vision, natural language processing, and speech recognition [1], and is now being widely applied to bioinformatic problems ranging from clinical image classification [2] to prediction of protein-DNA binding [3, 4]. A major driver of the success of deep learning has been the availability of standardized data sets such as ImageNet [5], which address three key needs: (i) fair apples-to-apples comparisons with existing algorithms, providing a reference point for the state of the art via a universal benchmark, (ii) at will assessment so that methods can be tried and tested rapidly and new results reported immediately—this has led to weekly publication of new machine learning algorithms—and (iii) access to pre-formatted data with the necessary inputs and outputs for supervised learning. While some bioinformatic applications enjoy this level of standardization [6], the central problem of protein structure prediction remains one without a standardized data set and benchmark. Availability of such a data set can spur new algorithmic developments in protein bioinformatics and lower the barrier to entry for researchers from the broader machine learning community.

Addressing the above needs for protein structure prediction necessitates a data set with several key features. First, sequence and structure data must be provided in a form readily usable by machine learning frameworks, standardizing the treatment of structural pathologies such as missing residues and fragments and non-contiguous polypeptide chains. Second, multiple

Correspondence: alquraishi@hms.harvard.edu
Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA

sequence alignments (MSAs) comprised of evolutionarily related proteins for every structure should be made available, given the central role that evolutionary information plays in modern protein structure prediction [7]. This is especially important as the generation of MSAs can be computationally demanding. Third, standardized training/validation/test splits that partition the data into subsets for fitting model parameters (training set), fitting model hyperparameters (validation set), and model assessment (test set) are needed to ensure consistency when training and assessing different learning algorithms [8]. Creating such splits can be straightforward for machine learning tasks involving images or speech, as data points from these modalities can be reasonably approximated as independent and identically distributed (IID). Natural protein sequences are far from IID however due to their underlying evolutionary relationships, a problem further exacerbated by the discrete nature of these sequences which can result in similar proteins having nearly identical numerical representations (this problem is avoided by e.g. images, as even small changes in lighting or viewing angle result in entirely different pixel-level representations). Consequently careful treatment of data splitting is required to ensure meaningful separation between subsets. Finally, multiple test objectives should ideally be provided to enable nuanced assessment of new methods, for example by testing varying levels of generalization capacity.

While a variety of protein structure databases do exist, none satisfy all the above requirements (Table 1). Repositories such as the Protein Data Bank (PDB) [9] provide raw protein structures, but require post-processing before they are usable by machine learning frameworks. Processed data sets such as CulledPDB [10] provide a more standardized preparation of protein structures, but lack evolutionary data such as MSAs. In fact, to our knowledge there is currently no public resource for high-quality MSAs suitable for protein structure prediction. One MSA repository does exist [11], but it appears out of date and is unsuitable for applications requiring deep homology searches [12]. The substantial computational cost associated with generating MSAs may explain this surprising absence.

With respect to standardized training/validation/test splits, the closest existing analogues are the biennial Critical Assessment of protein Structure Prediction (CASP) [13] and the continually running Continuous Automated Model EvaluatiOn (CAMEO) [14]. Both of these ongoing experiments provide an invaluable service for assessing prediction methods in a blind fashion, by presenting predictors with protein sequences whose structure has been solved but not yet made publicly available. Nonetheless, these experiments serve a different purpose from a standardized data split. CASP occurs once every two years, making it too infrequent for rapidly developing fields like machine learning. And while CASP can be thought to provide a training/test split based on the data available on the starting day of a given CASP assessment, it does not provide a validation set. Effective validation sets must mimic the generalization challenge presented to a trained model by the test set, by mirroring the distributional shift in data between the training and test sets; in effect, acting as a proxy for the test set. This is challenging to do for CASP proteins as they often contain novel protein folds occupying the twilight zone of sequence homology relative to PDB proteins (< 30% sequence identity [15]). Creating a matching validation set is thus non-trivial owing to the difficulty of detecting weak homology [16, 17].

Unlike CASP, CAMEO is continually running and thus can be used for assessment at any time. However, by virtue of its dynamic nature, CAMEO is difficult to use for apples-to-apples comparisons with an existing method unless both methods are participating simultaneously. CAMEO also focuses on proteins with known folds, making it less suitable for testing generalization to unknown parts of protein fold space.

To address these challenges and provide a community resource that promotes the application of machine learning to protein structure, we created ProteinNet.

Table 1 Summary of ProteinNet features relative to other database and repositories

Database	Structure	Sequence	PSSM/MSA	Clustering	Train/Val splits	Historical CASP reset	ML framework file format
PDB	Raw	✓	✗	Sequence	✗	✗	✗
CulledPDB	Processed	✓	✗	MSA	✗	✗	✗
HSSP	✗	✓	HSSP	✗	✗	✗	✗
ProteinNet	Processed	✓	JackHMMer	MSA	✓	✓	TensorFlow

Three existing databases are compared with ProteinNet in terms of available sequence, structure, and evolutionary profile information (PSSMs: position-specific scoring matrix; MSA: multiple sequence alignment), as well as standardized splits and tooling to facilitate machine learning (ML) applications. A ✓ indicates inclusion of a feature while a ✗ indicates exclusion. Structures can be raw or processed, with the latter indicating structure selection based on experimental quality metrics (e.g. R-factor) and annotation of structural pathologies (e.g. missing residues). PSSM/MSA indicates method used to derive evolutionary profiles. Note that HSSP is no longer widely used by the protein structure prediction community, while JackHMMer is one of the standard methods. Clustering can either be performed by sequence alignment or by exploiting MSAs to detect low sequence homology. The MSA approach used in ProteinNet can detect homology down to 10% sequence identity, which is not done by CulledPDB. Data splits segregating training and validation sets and resetting the historical record to reflect the state of prior CASPs is also unique to ProteinNet

ProteinNet provides pre-formatted input/output records comprising protein sequences, high-quality MSAs, and secondary and tertiary structures, as well as standardized data splits, including validation sets that emulate the generalization challenges presented by CASP proteins.

Methods

Design and approach

ProteinNet's design philosophy is simple: piggyback on the CASP series of assessments to create a corresponding series of data sets in which the test set is comprised of all structures released in a given CASP, and the training set is comprised of all protein structures and sequences (for building MSAs) publicly available prior to the start date of that CASP. A subset of the training data is set aside to create multiple validation sets at different sequence identity thresholds (relative to the training set), including < 10% to test generalizability to new protein folds comparable in difficulty to those encountered in CASP. Each ProteinNet data set effectively reverts the historical record to mimic the conditions of a prior CASP. We use CASP 7 through 12 (dating back to 2006) to create the corresponding ProteinNet 7 through 12. Our approach has three desirable properties.

First, by utilizing CASP structures for the test set, we leverage an objective third party's (the CASP organizers') selection of structures that meaningfully differ from the publicly accessible universe of PDB structures at a given moment in time. In particular, CASP organizers place prediction targets in two categories: template-based modeling (TBM) for proteins with clear structural homology to PDB entries, and free modeling (FM) for proteins containing novel folds unseen or difficult to detect in the PDB. This delineation provides an independent measure of difficulty useful for assessing models' ability to generalize to unseen parts of fold space. (CASP organizers occasionally include a third category, "TBM/FM" or "TBM hard", for structures of medium difficulty.)

Second, by utilizing multiple validation sets with varying levels of sequence identity, ProteinNet provides proxies for both TBM (20–40% seq. id.) and FM (< 10% seq. id.) CASP proteins. This enables optimization of model hyperparameters through monitoring of model generalization to proteins similar in difficulty to CASP TBM or FM proteins, potentiating the development of models focused exclusively on novel or known fold prediction. We note that this is distinct from merely having separate TBM and FM test proteins (first property), as test sets are only used for final model assessment and are thus unsuitable for hyperparameter optimization (the purpose of validation sets).

Third, by virtue of being the standard for assessing structure prediction methods, CASP enjoys the participation of all major predictors. It thereby provides a record of the accuracy of current and prior methods given available data at assessment time. Crucially, new methods trained and tested on ProteinNet demonstrate their performance on the same data splits as CASP-assessed methods, making them immediately comparable to state of the art methods on current and prior CASPs. This circumvents the catch-22 problem facing new benchmarks by providing immediate value to ProteinNet-trained models. Comparisons using older CASPs provide assessments with varying amounts of data, stressing algorithms in data rich and data poor regimes, a useful property when controlling for algorithmic vs. data-driven improvements, particularly in co-evolution-based methods.

Structures and sequences

All current PDB structures were downloaded using the mmCIF file format [18] then filtered by public release date so that ProteinNet training and validation sets only include entries publicly available prior to the start of their corresponding CASP assessment (Table 2). We exclude structures containing less than two residues or where > 90% of residues were not resolved, but otherwise retain virtually the entirety of the PDB. Mask records are generated for each structure to indicate which residues or fragments, if any, are missing, to facilitate processing by machine learning algorithms (e.g. by using a loss function that ignores missing residues). Multiple logical chains (in the same mmCIF file) that correspond to a single physical polypeptide chain are combined (with missing fragments noted), while physically distinct polypeptide chains are treated as separate structures. For chains with multiple models, only the first one is kept.

Sequences are derived directly from the mmCIF files. In instances where an amino acid is chemically modified or its identity is unknown, the most probable residue in its position-specific scoring matrix (PSSM) [19] is substituted (see next section for how PSSMs are derived). If a PSSM contains more than three adjacent residues with zero information content then its

Table 2 ProteinNet summary statistics

Data set	Cutoff date	Structures	Sequences
ProteinNet 7	2006/5/1	34,557	4,817,827
ProteinNet 8	2008/5/5	48,087	15,756,117
ProteinNet 9	2010/5/3	60,350	24,688,095
ProteinNet 10	2012/5/1	73,116	63,477,198
ProteinNet 11	2014/5/1	87,573	173,908,140
ProteinNet 12	2016/5/1	104,059	332,283,871

Cutoff dates for inclusion of sequence and structure data, based on the start of prior CASP experiments, are shown along with the number of sequences and structures in each ProteinNet set. Numbers are for non-redundant entries

corresponding sequence/structure is dropped, as we have found this to be a strong indicator that the sequence cannot be faithfully resolved.

In addition to full length PDB structures and sequences, single domain entries are created by extracting domain boundaries from ASTRAL [20] to enable training of both single and multi-domain models.

Multiple sequence alignments

Sequence databases for deriving MSAs were created by combining all protein sequences in UniParc [21] with metagenomic sequences from the Joint Genome Institute [22] and filtering to include only sequences available prior to CASP start dates (Table 2). JackHMMER [23] was then used to construct MSAs for every structure by searching the appropriate sequence database. Different MSAs were derived for the same structure if it occurred in multiple ProteinNets. JackHMMER was run with an e-value of 1e-10 (domain and full length) and five iterations. A fixed database size of 1e8 (option -Z) was used to ensure constant evolutionary distance when deriving MSAs (similar to using bit scores). Only perfectly redundant sequences (100% seq. id.) were removed from sequence databases to preserve fine- and coarse-grained sequence variation in resulting MSAs.

In addition to raw MSAs, PSSMs were derived using Easel [24] in a weighted fashion so that similar sequences collectively contributed less to PSSM probabilities than diverse sequences. Henikoff position-based weights (option -p) were used for this purpose.

Data splits and thinning

For each CASP cutoff date, we partition the full complement of (pre-CASP) structures and their associated MSAs into one training set and multiple validation sets, all non-overlapping (Fig. 1). Partitioning is done iteratively, by first clustering sequences at the 10% sequence identity level, randomly extracting 32 clusters, and then reclustering the remaining sequences at the next sequence identity level. Seven thresholds are used (10, 20, 30, 40, 50, 70, 90%) resulting in seven validation sets each comprising 32 clusters. While the selection of clusters is random, clusters larger than 100 members are not considered to minimize data loss. One representative exemplar is then selected from each cluster and the remaining cluster members are removed entirely (exemplar selection criteria is described at the end). Structures that remain outside of all validation clusters comprise the training set. The choice of 32 clusters per validation subset ensures sufficient sampling of proteins from every sequence identity level without removing an unnecessarily large number of proteins from the training set (up to 3200 proteins per validation subset).

Obtaining coherent clusters at < 20% sequence identity is difficult due to weak homology between individual sequences. To overcome this we perform comparisons using the previously derived MSAs instead of using ndividual sequences, as they provide greater sensitivity by incorporating evolutionary information (left inset in Fig. 1). First, sequences within MSAs are redundancy filtered to 90% seq. id. using HHsuite [25] to lower the computational load. We then carry out an all-against-all MSA-to-MSA comparison using HHblits [26] with one iteration and local alignment (option -loc). HHblits is necessary for this step as JackHMMER is unable to perform MSA-to-MSA searches, but the MSAs used are the original, JackHMMER-derived ones. Based on the HHblits alignment scores, we cluster MSAs using MMseqs2 [27] with the sought sequence identity threshold, an e-value threshold of 0.001, and clustering mode 1, which constructs a graph covering all sequences then finds remote homologs using transitive connections. We do not impose a minimal coverage requirement on sequence hits; this overestimates sequence identity, as short proteins can match subparts of longer ones. We use this approach to be maximally conservative in our construction of validation sets, to safeguard against accidental information leakage between training and validation sets.

Training sets are further processed to generate overlapping subsets that vary in sequence redundancy (at 30, 50, 70, 90, 95, and 100% seq. id.) which we call “thinnings”. For every thinning except 100% we cluster the training set by applying MMseqs2 directly to individual sequences (no MSAs) with the sought sequence identity threshold and a sequence coverage requirement of 80%. This requirement ensures that individual domains are not grouped with multi-domain proteins that contain them, as some models may seek to leverage single domain information. We do not utilize a coverage requirement for the validation set to prevent information leakage, but it is not a concern for the training set. For the 100% thinning every set of identical sequences is used to form a cluster. After clusters are formed, a single exemplar is selected from each, and all remaining cluster members are removed.

We use the same exemplar selection criteria for validation and training sets. As a rule, we avoid selecting exemplars near cluster boundaries, as two boundary sequences in different clusters may be closer in sequence space than the sought sequence identity threshold (right inset in Fig. 1). To ensure this we always pick exemplars near the cluster center. By default, MMseqs2 returns an exemplar which is centermost in the cluster without incorporating other, potentially useful criteria such as structure quality. To incorporate these criteria, we use the MMseqs2 exemplar as bait to form a new cluster of sequences that are 95% identical to the exemplar and

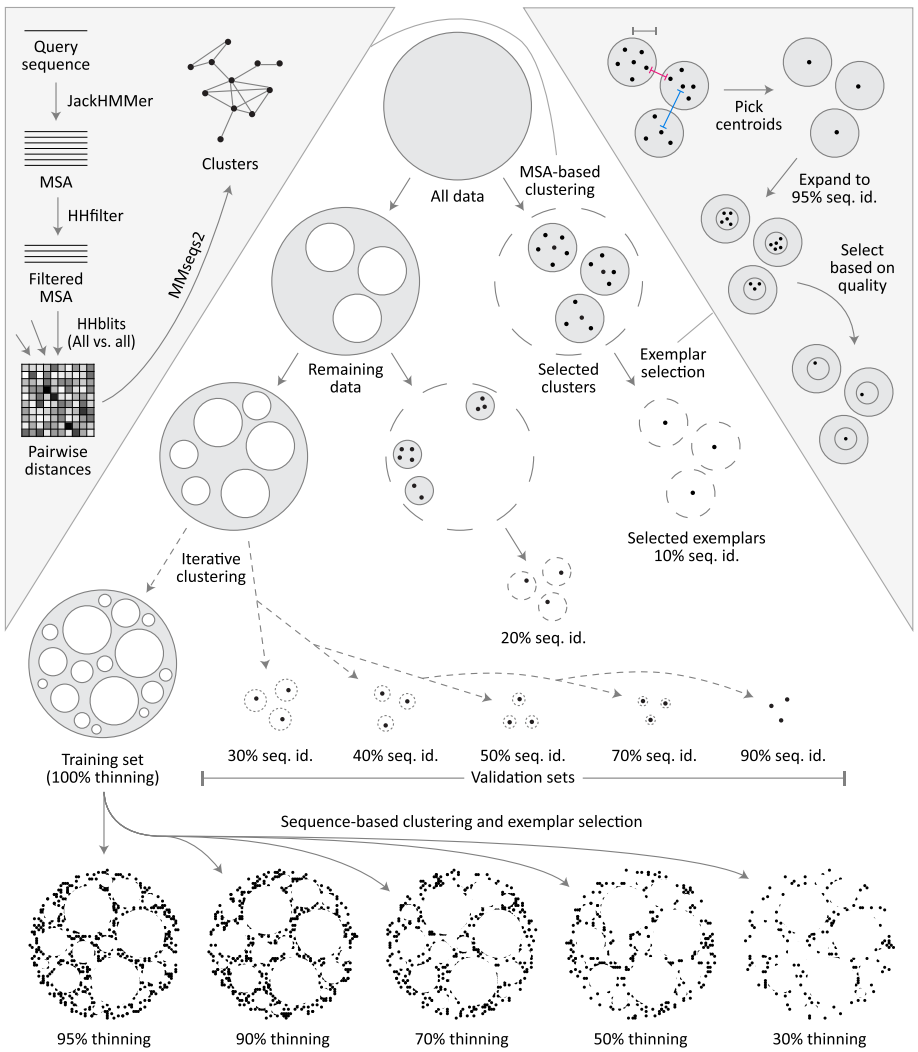

Fig. 1 ProteinNet construction pipeline. For each ProteinNet, all proteins with PDB structures available prior to the start of its corresponding CASP ("All data", top circle) are clustered using an MSA-based clustering technique (left inset) to yield large clusters where intra-cluster sequence identity is as low as 10%. One exemplar from each cluster is then selected (right inset) to yield the 10% seq. id. validation set. This process is iteratively repeated, by reclustering the data remaining outside of all initial clusters to yield validation sets of higher sequence identity (20–90%). Once the final validation set is extracted, all remaining data is used to form the training set. Based on this set ("100% thinning"), filtered training sets are created at lower sequence identity thresholds to provide coarser sampling of sequence space. Left inset: Each protein sequence is queried against a large sequence database (filtered to only include sequences publicly available prior to the beginning of the corresponding CASP) using JackHMMer to create an MSA that is subsequently filtered to 90% seq. id. HHblits is then used to perform an all-against-all sequence alignment of MSAs. Finally, alignment distances are fed to MMseqs2 to cluster their corresponding sequences. Right inset: The center-most protein of each cluster is chosen to ensure that the desired sequence identity constraints are satisfied, as proteins near cluster boundaries may be closer than the pre-specified radius of each cluster (pink vs. gray measuring tapes), while the distances between cluster centroids must satisfy the sequence identity constraints (blue measuring tape). The centroids are then used to form tight clusters of 95% seq. id. that are intersected with the original clusters to yield candidate exemplars ranked by multiple quality metrics (see main text). The top-ranked candidate is picked as the exemplar protein of each cluster

cover 90% of its length, yielding a tight cluster that is highly sequence-similar but with potentially better structural characteristics. From the intersection of the original MMseqs2-derived cluster and the new one, we then pick the structure that optimizes the following criteria, in order: structure quality (defined as 1 / resolution - R-value, the same criterion used by the PDB), date of release (newer is better), and length (longer is better).

File formats and availability

All sequences, structures, MSAs, and PSSMs have been made available for download individually in standard file formats. In addition, ProteinNet records integrating

sequence, structure (secondary and tertiary), and PSSMs in a unified format are available as human-readable text files and as binary TensorFlow records [28]. We provide Python code for parsing these records directly into TensorFlow to facilitate their use in machine learning applications.

Utility and discussion

We applied the ProteinNet construction pipeline to CASP 7 through 12, resulting in six data sets ranging in size from 34,557 to 104,059 structures (Table 2). We observe a generally linear increase in the number of training structures, across all thinnings, over this time period (Fig. 2a), consistent with the PDB's sequence bias remaining constant. The growth in sequence data on the other hand is exponential (Table 2), much of which driven by metagenomic databases comprised largely of prokaryotic genes. Since unknown prokaryotic genes are less likely to be crystallized, they are not well presented among CASP targets [22]. Nonetheless, the growth in sequence databases has resulted in higher quality MSAs in later CASPs, as measured by the number of sequences in alignments (Fig. 3). The overall number of CASP test structures is roughly constant, although the proportion of FM targets has increased (Fig. 4), likely reflecting the end of the Structural Genomics Initiative [29] which crystallized a large number of proteins of known folds.

Examining sequence length, we observe that CASP structures have on average grown in length (Fig. 4), and similarly for the PDB (Fig. 2 b, c), although the vast majority of structures remain shorter than 1000 residues. This trend is likely to accelerate with increased use of CryoEM [30] methods which have made multi-domain proteins more amenable to structural characterization.

To assess the suitability of ProteinNet validation sets to serve as proxies for CASP targets, we computed the distance, measured by sequence identity, of every entry in the ProteinNet validation and test sets to its closest entry in the training set. Because sequence identity is difficult to detect in low homology regions (< 30% seq. id.), we first performed an all-against-all alignment using MSA-MSA searches, similar to our pipeline for clustering, and then computed sequence identity using the resulting matches. We used an e-value threshold of 0.001 to ensure genuine hits, but otherwise imposed no additional constraints. Figure 5 summarizes the results. As expected, FM targets across all CASPs show no detectable sequence homology to their corresponding ProteinNet training sets. Importantly, the < 10% seq. id. validation sets of all ProteinNets show no detectable homology to the training sets either, indicating that they can act as proxies of CASP FM targets. TBM targets roughly exhibit between 10 and 30% seq. id. to the ProteinNet training sets, similar to the < 20, < 30, and < 40% seq. id. validation sets, confirming that they can act as proxies of CASP TBM targets. We conclude that the appropriate ProteinNet validation set can be used to optimize models whose goal is to generalize to protein folds similar in difficulty to CASP FM and TBM targets. ProteinNet validation sets with higher sequence identity,

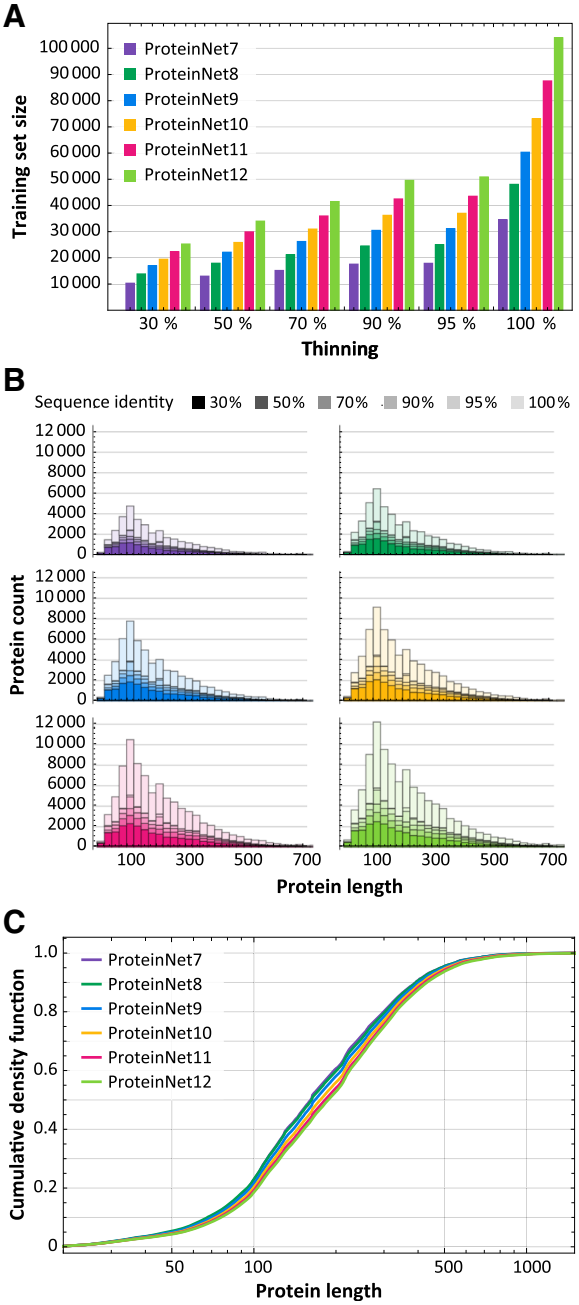

Fig. 2 Statistics of ProteinNet data sets. **a** Number of proteins in ProteinNet training sets for different thinnings (30–100% seq. id.) **(b)** Protein length distributions for ProteinNet training sets. **c** Cumulative density function of protein length distribution for 100% thinnings of ProteinNet training sets

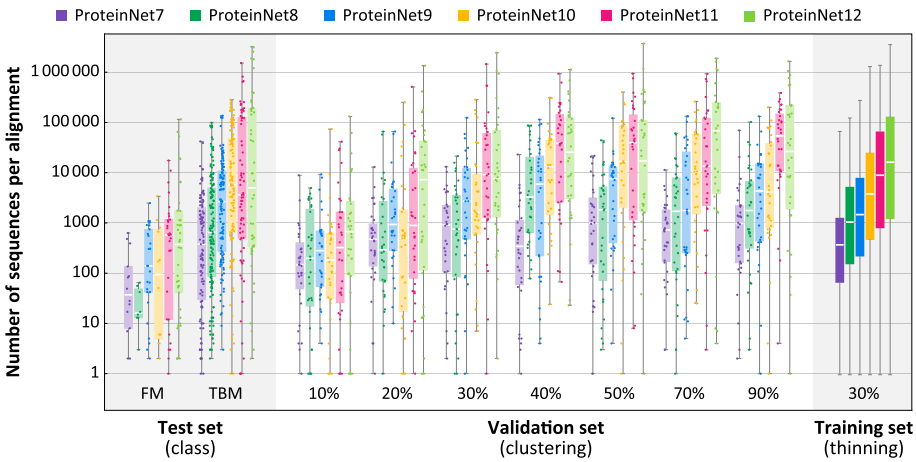

Fig. 3 Alignment size as a function of ProteinNet subset. Box and whisker charts depict the distribution of number of sequences per MSA for ProteinNet training (30% thinning), validation, and test sets. Individual data points for training sets are not shown due to their large size

i.e. > 50%, are potentially useful for optimizing models focused on predicting changes to known protein structures, such as those induced by mutations.

We next sought to assess how growth in the number of PDB structures changes the difficulty of CASP TBM targets. For every CASP test set, we repeated the previous analysis using older ProteinNet training sets. E.g., for CASP 11, we compared its TBM set against ProteinNet 7–11 training sets. Figure 6 summarizes the results. As expected, earlier ProteinNet training sets show greater distance from the TBM sets, particularly for older CASPs, with a general loss of ~ 2–3% seq. id. points per CASP (i.e. two years). This type of retroactive analysis may be used to assess an algorithm's sensitivity to the amount of available data, with Fig. 6 providing a characterization of the relative difficulty of different CASP targets when using different ProteinNets for training (raw distance data at the single protein level is available at the ProteinNet repository). We did not perform this analysis for FM targets since even the most up to date ProteinNet training sets (for a given CASP) do not show any detectable homology, thereby precluding older training sets from showing further homology.

Conclusions

Standardized data sets have unlocked progress in myriad areas of machine learning, and biological problems are no exception. ProteinNet represents a community resource for bioinformaticists and machine learning researchers who seek to test new algorithms in a manner consistent with state of the art blind assessment. It lowers the barrier to entry for the field by aggregating the relevant data modalities in a single file format, and by eliminating the upfront computational cost required for creating high-quality MSAs. Collectively, the generation of all MSAs and PSSMs in ProteinNet 7–12

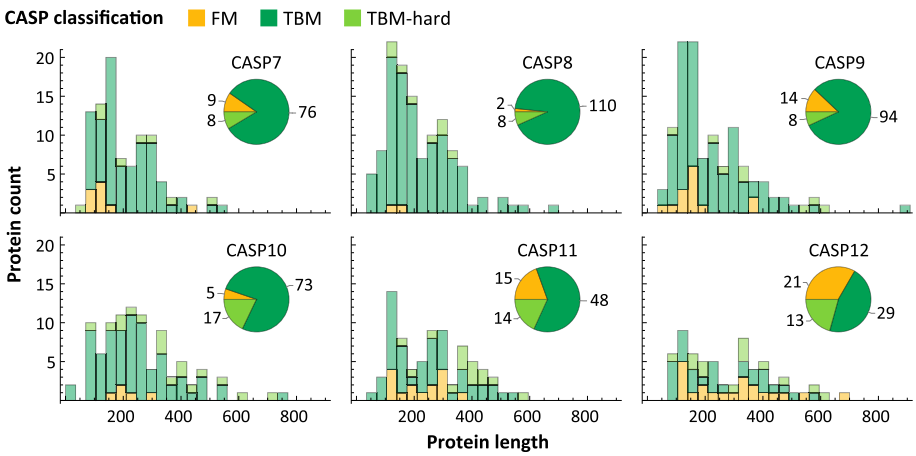

Fig. 4 Statistics of CASP data sets. Length distribution of proteins in CASP 7 through 12, broken down by difficulty class. Pie charts show the number of proteins per difficulty class

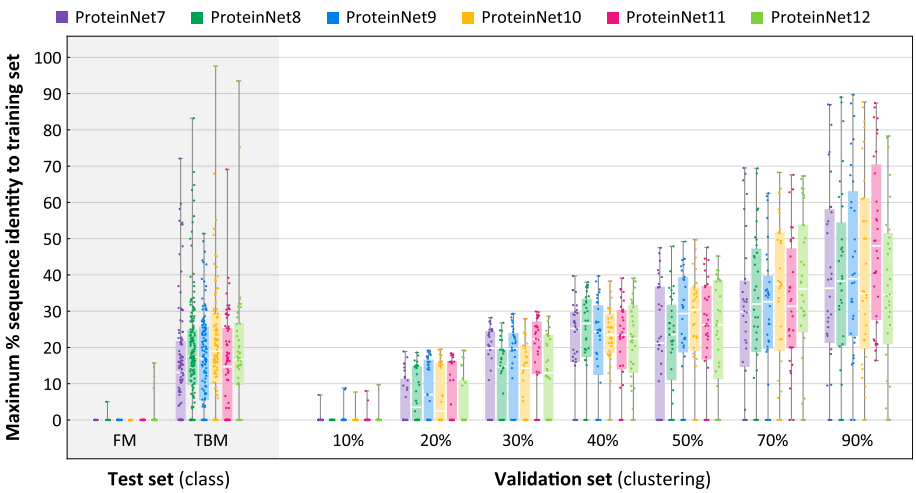

Fig. 5 Distributions of maximum % sequence identity to training sets. Box and whisker charts depict the distribution of maximum % sequence identity, with respect to the training set, of all entries in a given ProteinNet validation or test set. The FM test sets and 10% seq. id. validation sets show a median value of 0% seq. id. to the training set

consumed over 3 million compute hours, a one-time investment whose benefits can now be shared by the entire community of researchers. Perhaps most crucially, ProteinNet provides validation sets that provide a reliable assessment of model generalizability, ensuring that progress can be meaningfully ascertained while training models.

Beyond protein structure prediction, ProteinNet can serve as a data set for a number of important problems. ProteinNet prescribes no intrinsic preference for which data modalities should serve as inputs and which should serve as outputs. A protein design algorithm can hypothetically be trained by using structures as inputs, and the sequences of their associated MSAs as outputs. Alternatively, an algorithm for predicting the effects of mutant variants can use the sequence and structure of one protein as input, and output the structures of proteins with similar sequences as predictions.

More broadly, the advent of deep learning methods and automatic differentiation frameworks like TensorFlow and PyTorch [31] makes it possible to build bespoke models of biological phenomena. In the machine learning community, this has spurred the development of so-called multi-task learning problems in which multiple output modalities are simultaneously predicted from a given input, as well as auxiliary losses

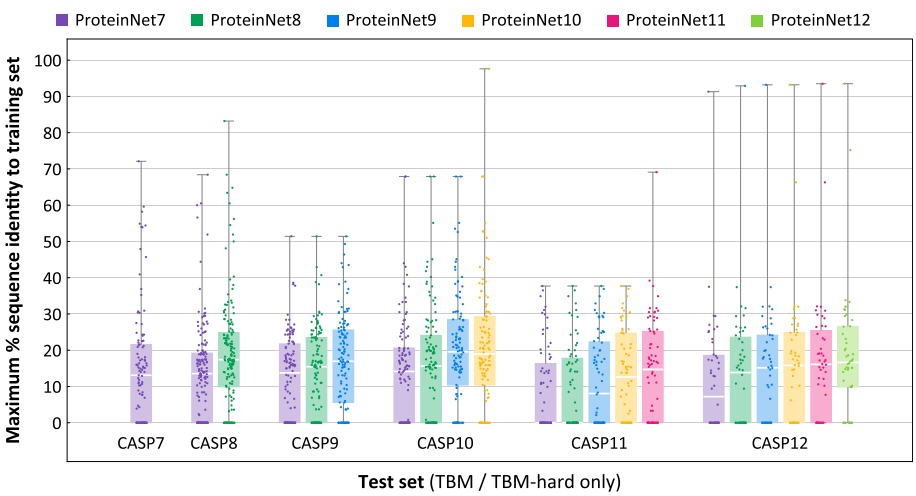

Fig. 6 Distributions of maximum % sequence identity of CASP entries with respect to prior training sets. Box and whisker charts depict the distribution of maximum % sequence identity, with respect to a training set, of all TBM/TBM-hard entries in a given ProteinNet test set (CASP set). Comparisons are made for each ProteinNet test set with respect to its corresponding and prior training sets, e.g. for CASP 11 with respect to ProteinNet 7–11 training sets. Color indicates training set used

in which a core objective function is augmented with additional output signals that can help train a more robustly generalizing model. In many gene- or protein-related learning tasks, protein structure is one such broadly useful output signal that can augment a supervised learning problem, e.g. the prediction of the DNA binding affinity of a transcription factor, with information that is proximal to the desired task. ProteinNet should help facilitate such applications, along with the development of end-to-end differentiable models of protein structure that can be directly fused to other learning problems [32]. As the quality of protein structure prediction algorithms continues to improve, we believe that structural information will get increasingly integrated within a wide swath of computational models.

Abbreviations

CAMEO: Continuous Automated Model EvaluatiOn; CASP: Critical Assessment of protein Structure Prediction; FM: Free modeling; IID: Independent and identically distributed; ML: Machine learning; MSA: Multiple sequence alignment; PDB: Protein Databank; PSSM: Position-specific scoring matrix; TBM: Template-based modeling

Acknowledgements

We thank Peter Sorger for his mentorship and support, and Uraib Aboudi for her editorial comments and helpful discussions. We also thank Martin Steinegger and Milot Mirdita for their help with using the HHblits and MMseqs2 packages, Sergey Ovchinnikov for help with metagenomics sequences, Andriy Kryshtafovych for his help with CASP structures, Sean Eddy for his help with using the JackHMMer package, and Raffaele Potami, Amir Karger, and Kristina Holton for their help with using HPC resources at Harvard Medical School.

Authors' contributions

MA collected the raw data and designed the workflow for creating the data set. MA wrote and approved the manuscript.

Funding

This work has been supported by NIGMS grant P50GM107618 and NCI grant U54-CA225088. NIGMS and NCI were not involved in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the ProteinNet GitHub repository, https://github.com/aqlaboratory/proteinnet.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 18 March 2019 Accepted: 5 June 2019
Published online: 11 June 2019

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
2. Ting DSW, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med. 2018;24(5):539.
3. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
4. Ching T, Himmelstein DS, Beaulieu-Jones Brett K, Kalinin Alexandr A, Do Brian T, Way Gregory P, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):20170387.
5. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52.
6. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol. 2018;36:391–2.
7. de Oliveira S, Deane C. Co-evolution techniques are reshaping the way we do structural bioinformatics. F1000Research [Internet]. 2017 [cited 2019 Jan 22];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5531156/
8. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: The MIT Press; 2016. p. 800.
9. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The protein data Bank: a computer-based archival file for macromolecular structures. J Mol Biol. 1977;112(3):535–42.
10. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589–91.
11. Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39(Database issue):D411–9.
12. Rost B. Twilight zone of protein sequence alignments. Protein Eng Des Sel. 1999 Feb 1;12(2):85–94.
13. John M, Krzysztof F, Andriy K, Torsten S, Anna T. Critical assessment of methods of protein structure prediction (CASP)—round XII. Proteins Struct Funct Bioinforma. 2018;86(S1):7–15.
14. Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, et al. Continuous automated model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins Struct Funct Bioinforma. 2018;86(S1):387–98.
15. Khor BY, Tye GJ, Lim TS, Choong YS. General overview on structure prediction of twilight-zone proteins. Theor Biol Med Model [Internet]. 2015 Sep 4 [cited 2019 Jan 22];12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4559291/
16. Habermann BH. Oh Brother, Where Art Thou? Finding Orthologs in the Twilight and Midnight Zones of Sequence Similarity. In: Pontarotti P, editor. Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods [Internet]. Cham: Springer International Publishing; 2016 [cited 2019 Jan 22]. p. 393–419. Available from: https://doi.org/10.1007/978-3-319-41324-2_22
17. Chen J, Guo M, Wang X, Liu B. A comprehensive review and comparison of different computational methods for protein remote homology detection. Brief Bioinform. 2018 Mar 1;19(2):231–44.
18. Westbrook JD, Fitzgerald PMD. The PDB format, mmCIF formats, and other data formats. In: Structural bioinformatics [internet]. John Wiley & Sons, Ltd; 2005 [cited 2019 Jan 24]. p. 159–179. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471721204.ch8.
19. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16–23.
20. Fox NK, Brenner SE, Chandonia J-M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014;42(D1):D304–9.
21. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2018; 46(5):2699–2699.
22. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017;355(6322):294–8.
23. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7(10):e1002195.
24. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46(W1):W200–4.
25. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005 Apr 1;21(7):951–60.
26. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012 Feb;9(2):173–5.
27. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Oct 16;35:1026–8.
28. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) [Internet]. 2016

[cited 2019 Jan 22]. p. 265–283. Available from: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
29. Chandonia J-M, Brenner SE. The impact of structural genomics: expectations and outcomes. Science. 2006;311(5759):347–51.
30. Callaway E. The revolution will not be crystallized: a new method sweeps through structural biology. Nature. 2015;525(7568):172–4.
31. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017 [cited 2019 Jan 22]; Available from: https://openreview.net/forum?id=BJJsrmfCZ
32. AlQuraishi M. End-to-end differentiable learning of protein structure. Cell Syst. 2019 Apr 24;8(4):292–301.e3.

Publisher's Note