Pre-processing annotated homologous regions in protein sequences concerning machine-learning applications

Nawar Malhis
nmalhis@msl.ubc.ca
Michael Smith Laboratories, University of British Columbia
Vancouver, BC V6T 1Z4, Canada

Abstract

While annotated protein sequences are widely used in machine learning applications, pre-processing these sequences regarding homology is mainly limited to clustering complete sequences based on global alignment without considering their annotations. Here, I am introducing new tools that identify all possible local homologies between annotated sequences within the same or across two datasets and then resolve these homologies.

Availability: `orca.msl.ubc.ca/nmshare/ham.tar.gz`

1 Introduction

Machine learning tools are widely used to predict different functional features of protein sequences, such as Antimicrobial peptides (AMPs) [1], antigen-antibody affinity [2], post-translational modification (PTM) sites [3], and protein-disordered binding sites [4][5]. Data of protein sequences annotated with the targeted feature are collected and then divided into a training and a testing dataset. The training dataset is used to train some machine learning models, and then the testing dataset is used to evaluate the trained model's predictions. Two important issues need to be considered. **First**, we must ensure no data leakage between the training and testing datasets. In theory, training and testing datasets should be separated at random. However, this random splitting can lead to data leakage in the case of protein sequences, as protein sequences tend to include a high percentage of homologous sequences. It is a common practice to cluster protein sequences at some identity cut-off and then split the clusters randomly so that homologous sequences in each cluster are allocated to either the training or the testing dataset. The clustering identity cut-off is often chosen based on the targeted functional feature. If this feature can be located within short regions, this cut-off must be low enough to guarantee, to the degree possible, homology-free splitting concerning the critical targeted feature. **Second**, we should not allow for annotation conflict. As homologous sites are likely to have similar molecular functions, and most importantly, they will be predicted alike, having homologous regions with conflicting annotations within the training dataset can negatively impact the training process. Similarly, conflicting annotations in the test data can yield a sub-optimal evaluation.

In this paper, I'm introducing two tools that identify homology in annotated protein sequences and solve the abovementioned issues. The first tool, HAM (Measuring Annotated Homology), is to solve the data leakage issue. HAM identifies homologous regions between annotated protein sequences in different datasets, highlights cross annotations, and then masks these regions in one of the two datasets. The

second tool, HAC (Homology Annotation Conflict), identifies local homology between sequences within the same dataset, highlights conflicts in annotations, and resolves these conflicts.

2 Results

I examined three datasets, each divided into training and testing, and used to train and test predictors of intrinsically disordered protein binding sites. The first dataset (D2008) is a popular dataset assembled in 2008 [6] and used to train and test several MoRF predictors, including MoRFpred [6], MFSPSSMpred [7], MoRFchibi [8], MoRFchibi_web [9], fMoRFpred [10], MoRFchibi_light [11], OPAL [12], MoRFmlp [13], and MoRFcnn [14]. The second dataset (D2021) was collected in 2021 to train and test the DeepDISObind IDR binding sites [15], and the third (D2022) was collected in 2022 and used to train and test CLIP, which predicts linear interacting peptides (LIP) [16].

The authors of MoRFpred assembled 840 sequences with IDR protein-binding sites, known as molecular recognition features, or MoRFs, that undergo a disorder-to-order transition upon binding to protein targets (D2008). They clustered these MoRFs annotated sequences at 30% identity using CD-Hit [17] into 427 clusters. Then, they randomly divided them into a training set with 421 sequences and a testing set with 419 sequences. The authors of DeepDISOBind collected 750 protein sequences from the DisProt [18] database with high-quality IDR binding site annotations that bind to either protein, DNA, or RNA targets (D2021). Here, I'm only considering IDR sites that bind to protein targets. Like D2008, sequences in D2021 are clustered at 30% identity into 650 clusters. Then, these clusters are divided into a training set with 356 sequences and a testing set with 394 sequences. Compared to D2008, D2021's higher number of clusters, despite its smaller number of sequences, implies a lower degree of homology between these sequences. The authors of CLIP started with more than 1,800 sequences annotated with Linear Interacting Peptides collected from the MobiDB 3.0 LIP database [19]. First, they used CD-Hit to cluster these sequences at 100% identity cut-off and retained the longest sequence in each cluster after transferring LIP annotations from other sequences in its cluster to it. Then, they clustered these sequences at 25% identity cut-off into 1,440 clusters. They selected the longest sequence of each cluster and discarded the remaining sequences. Finally, they randomly divided these sequences into a training set with 1,000 sequences and a testing set with 440 sequences.

In short, when D2008 was assembled, the available known MoRFs were very limited; thus, to retain a reasonable overall dataset size, the authors of MoRFpred did not filter out homologous sequences. The D2021 included 750 protein sequences with high-quality binding IDR annotations and lower homology. The MobiDB LIP database collected annotations from several databases, resulting in a much larger overall sequence count, enabling CLIP authors to filter out all homologous sequences at 25% identity between the training and testing dataset, as well as within each of these dataset (D2022).

Even though the clustering approach described above was used to separate each of the three datasets into training and testing, using HAM, I was able to identify important data leaks between the training and testing sequences in each. **Table 1** lists the overall annotated sequences that overlap between the training and testing sequences in the three datasets. Note that these numbers between Dataset 1 and Dataset 2 need not be consistent with those between Dataset 2 and Dataset 1, as one residue in one dataset can be homologous to more than one residue in the other dataset, see **Figure 1**.

Table 1: The observed homology overlap between the training and testing datasets annotated sequences using HAM with default parameters. Identity cut-off is the clustering identity used to separate the sequences in these datasets. Homologous residue counts in Dataset 1 are presented. Total is the total count of homologous residues. (A1 to A2) is the number of residues annotated with A1 in Dataset 1 that are homologous to residues annotated with A2 in Dataset 2.

Dataset 1	Dataset 2	Identity cut-off	Total	0 to 0	0 to 1	1 to 0	1 to 1
TR2008	TS2008	30%	46	37	2	1	6
TS2008	TR2008	30%	76	61	1	3	11
TR2021	TS2021	30%	761	739	11	11	0
TS2021	TR2021	30%	702	680	11	11	0
TR2022	TS2022	25%	375	368	1	6	0
TS2022	TR2022	25%	375	368	6	1	0

TR2008:

```
>Q5XLP6_1_32            MASWSHPQFEKGAETAVPNSRVSKKKTAAPKE
Annotation              00001111110000000000000000000000
Disordered              11111111111111111111111111100000
```

TS2008:

```
>Q8CH90_725_743 LRETSLVPRGSWSHPQFEK
Annotation      0000000000000111111
Disordered      0000000000000000000

>Q8CJD2_686_704 ASGVDLVPRGSWSHPQFEK
Annotation      0000000000011111111
Disordered      0000000000000000000

Alignment                  *********
```

Figure 1: Despite the training TR2008 and the testing TS2008 datasets were separated at 30% identity cut-off, the HAM software identified a nine residues region with 100% identity shared between the N-terminus of the training protein sequence Q5XLP6 (1kl3_E) and the C-terminus of two protein sequences in the TS2008 testing dataset, Q8CH90 (1kl3_F) and Q8CJD2 (1rsu_P). Six of these nine residues in the training sequence Q5XLP6 are annotated as MoRF residues, and six and eight are annotated as MoRFs in each of the testing sequences Q8CH90 and Q8CJD2.

Figure 2 presents a case where the homologous residues between training and testing datasets are annotated differently. Although the alignment of these two protein sequences results in an identity of 5.3%, the HAM software identified an eleven residues region shared between TR2021 protein P63315 (90-100) and TS2021 protein P16471 (286-296) with an identity 9/11. These 11 residues are annotated as IDR-binding residues in P16471 and non-binding in P63315. However, the MoRFchibi_web protein binding prediction scores suggest that P63315 residues 90-100 are IDR-binding residues, see **Figure 2**.

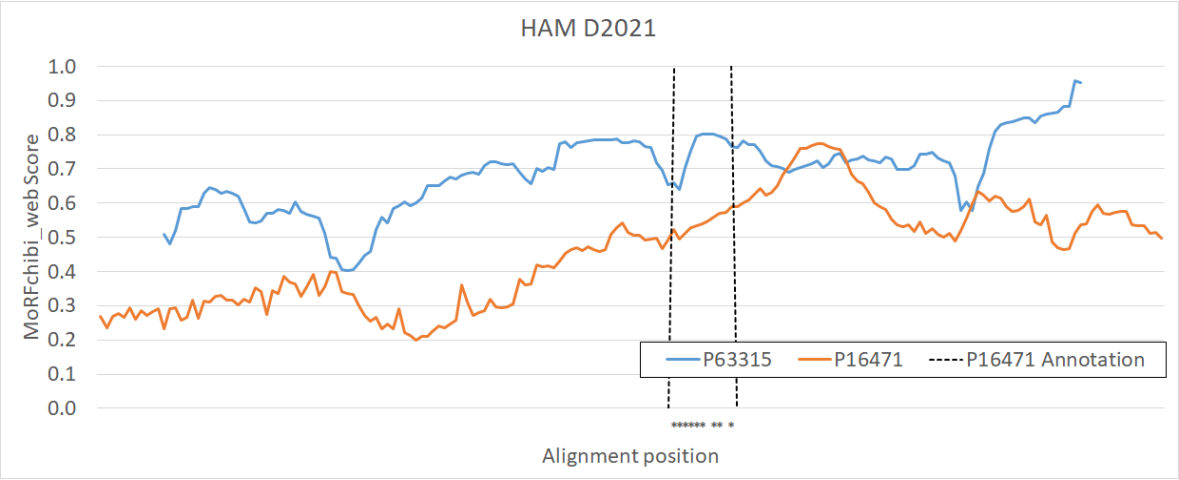

Figure 2: The MoRFchibi_web scores for P63315 and part of P16471, from 186 to 371. Only the region with residues marked with an '*' are aligned between these sequences. This region was not identified as a protein-binding region on the training side (P63315); on the testing side (P16471), it is, and the MoRFchibi_web score predicted the training region to be a protein-binding MoRF.

HAM identified the C-terminus 98 residues in protein Q13351 (KLF1_HUMAN size 362) in the training dataset TR2022 to be homologous to their counterpart residues in protein Q60793 (KLF4_MOUSE size 483) in the testing dataset TS2022 with an identity of 83.7% and both regions are annotated as not protein-binding, which leads to data leakage. HAM identifies such homologous regions and allows them to be masked in one of the datasets.

In contrast to HAM, the Homology Annotation Conflict software, HAC, identifies homologous regions, including those with conflicting annotations within the same dataset. **Table 2** presents the annotated homology between sequences in the six datasets identified by the HAC software using the default parameters.

Table 2: Using the HAC software with default parameters, the annotated homology within each of the six datasets. The Identity cut-off is the clustering cut-off used, if any. Total is the total homologous residues. (A1 to A2) is the number of residues annotated with A1 that are homologous to residues annotated with A2.

Dataset	Identity cut-off	Total	0 to 0	0 to 1	1 to 0	1 to 1
TR2008	None	89,309	84,278	1,652	1,092	2,287
TS2008	None	113,986	109,404	2,064	1,102	1,416
TR2021	None	3,938	2,840	353	353	392
TS2021	None	4,464	3,849	302	313	0
TR2022	25%	1,125	1,067	13	13	32
TS2022	25%	310	310	0	0	0

Figure 3 presents annotation conflict in a region of 13 residues shared between three Large ribosomal subunit protein sequences in the TS2021 dataset: P05318 residues 94 to 106, P05387 residues 103 to 115, and P23632 residues 95 to 107. This homologous region exists within a larger IDR segment annotated in the UniProt database [20] according to the sequence analysis SAM: MobiDB-lite [19] in all three protein sequences. **Figure 4** presents seven homologous sequences with annotation conflicts observed by HAC in the TR2008 dataset and HAC's three joint annotation options that can be used to resolve this conflict.

```
>P05318_84_106  EAEKEEEEAKEESDDDMGFGLFD
Annotation      0000000000000000000000
Disordered      1111111111111111111111
>P05387_93_115  EKKDEKKEESEESDDDMGFGLFD
Annotation      0000000001111111111111
Disordered      1111111111111111111111
>P23632_85_107  GAAAPAAAAEEEEDDDMGFGLFD
Annotation      0000000000111111111111
Disordered      1111111111111111111111
Matching AAs              ** **********
```

Figure 3: The aligned disordered C-terminus of three Large ribosomal subunit proteins from three TS2021 different species: Yeast (P05318), Human (P05387), and Trypanosoma cruzi (P23632). This alignment includes a 13 residues homology with an identity between each two sequences of at least 12/13. This region is annotated as protein-binding in P05387 and P23632 and non-binding in P05318.

```
>P28505|1a3b_I      ITYTDCTESGQDLCLCEGSDVCGKGNKCILGSNGEENQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000011111111110
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111111

>P28504|1a4w_I      ITYTDCTESGQDLCLCEGSNVCGKGNKCILGSNGEENQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000011111111100
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111111

>P28505|1d4p_H      ITYTDCTESGQDLCLCEGSDVCGKGNKCILGSNGEENQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000011111111111
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111111

>P28510|1fpc_I      VVYTDCTESGQNLCLCQDSNVCGQGNKCILGSNGEKNQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000011111111110
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111111

>P28503|1ghv_I      ITYTDCTESGQNLCLCEGSNVCGNGNKCKLGSDGEENQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000011111111111
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111111

>P28509|1lhf_I      VVYTDCTESGQNLCLCEDSNVCGQGNKCILGSNGEKNQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000000111111100000
H0                  00000000000000000000000000000000000000000000000000.........00
H1                  ..................................................111111111111111

>P28507|1vit_I      VVYTDCTESGQNLCLCEDSNVCGEGNKCILGSNGEKNQCVTGEGTPKPQSHNDGDFEEIPEEYLQ
Annotation          00000000000000000000000000000000000000000000000011111111111111
H0                  00000000000000000000000000000000000000000000000000.....00000
H1                  ..................................................111111111111

                    000000000111111111122222222223333333333444444444455555555556666666
AA Index            123456789012345678901234567890123456789012345678901234567890123456789012345
Alignment             ********* ****  * *** **** *** ** ****************************

Conflicting To 1    00000000000000000000000000000000000000000000000011111111111111
Conflicting To 0    00000000000000000000000000000000000000000000000000011111100000
Conflicting To Mask 000000000000000000000000000000000000000000000000----111111-----
```

Figure 4: Seven aligned TR2008 annotated equal-size sequences with Conflicting annotations. H0 is the '0'

annotations observed in homologous sequences. H1 is the '1' annotations observed in homologous sequences. The last three lines represent the annotation resolutions provided by HAC: Conflicting to '1' reannotates conflicting annotation to '1'. Conflicting to '0' reannotates conflicting annotation to '0'. And Conflicting to mask masking conflicting reannotates, i.e., '-'.

HAC also identified long stretches of homologous annotated sequences. In the TS2021 dataset, HAC identified P05204 (a 90 amino acid Non-histone Homo-sapiens chromosomal protein HMG-17) and P02313 (a 90 amino acid Non-histone Bos-taurus chromosomal protein HMG-17) as homologous with an identity of 97.8% and an annotation conflict as the entire P05204 is annotated as protein-binding while the P02313 as non-binding. Both proteins are declared disordered in the UniProt database based on MobiDB-lite automatic assertion.

Another example of conflicting annotations HAC identified in TR2021 is in regions 1 to 77 in P0ABD8 and P0ABE0. Both proteins are Biotin carboxyl carrier proteins of acetyl-CoA carboxylase in E-coli, strain K12 vs O157:H7, with 156 residues each and 100% identity. Residues 1-77 are annotated as protein-binding in P0ABD8, while the entire P0ABE0 is annotated as non-binding.

3 Methods

This software includes four Python tools: `ham.py`, `ham_mask_homology.py`, `hac.py`, and `hac_resolve_conflicts.py`. Both `ham.py` and `hac.py` utilize `blastp` with "`-outfmt 6`" and "`-word_size 3`" to identify all possible local alignments between a query sequence and a database. For `ham.py`, the database is created from one (the larger) input dataset, and the other dataset sequences are used for queries. For `hac.py`, the input dataset is used to create the database and query sequences, where the query sequence matching itself is ignored. In both tools, all candidate matches with no gaps that are longer than some minimum threshold (default is ten amino acids), and an identity higher than a minimum identity cut-off (default 80%) are considered.

`ham.py` generates three output files: one for detailed one-to-one residue alignment matches and two annotated fasta files, one for each input file, which includes extra annotations representing the annotations of homolog sequences. `ham_mask_homology.py` reads all ham-generated annotated fasta for the input dataset and generates an output annotated fasta file with all homologous residues masked, i.e., annotated by a (-).

`hac.py` generates two output files: a detailed one-to-one residue alignment matches and the input file with extra annotations representing the annotations of homolog sequences. `hac_resolve_conflicts.py` reads the annotated fasta file generated by `hac.py` for the input file and resolves the observed annotation conflicts by reannotating residues with conflicting annotations by a '1', '0', or by masking them with a '-', see **Figure 4**.

References

1. Lawrence TJ, Carper DL, Spangler MK, Carrell AA, Rush TA, Minter SJ, Weston DJ, Labbé JL. amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. Bioinformatics. 2021 Aug 4;37(14):2058-2060. doi: 10.1093/bioinformatics/btaa917. PMID: 33135060.
2. Yuan Y, Chen Q, Mao J, Li G, Pan X. DG-Affinity: predicting antigen-antibody affinity with language models from sequences. BMC Bioinformatics. 2023 Nov 13;24(1):430. doi: 10.1186/s12859-023-05562-z. PMID: 37957563; PMCID: PMC10644518.
3. Piovesan D, Hatos A, Minervini G, Quaglia F, Monzon AM, Tosatto SCE. Assessing predictors for new post-translational modification sites: A case study on hydroxylation. PLoS Comput Biol. 2020 Jun 22;16(6):e1007967. doi: 10.1371/journal.pcbi.1007967. PMID: 32569263; PMCID: PMC7332089.
4. Basu S, Kihara D, Kurgan L. Computational prediction of disordered binding regions. Comput Struct Biotechnol J. 2023 Feb 10;21:1487-1497. doi: 10.1016/j.csbj.2023.02.018. PMID: 36851914; PMCID: PMC9957716.
5. Malhis, N, Gsponer, J. Computational Prediction of Linear Interacting Peptides. Springer Natures. Methods in Molecular Biology book, "Prediction of Protein Secondary Structure" second edition, chapter 14. In press. (2024).
6. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012 Jun 15;28(12):i75-83. doi: 10.1093/bioinformatics/bts209. PMID: 22689782; PMCID: PMC3371841.
7. Fang C, Noguchi T, Tominaga D, Yamana H. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics. 2013 Oct 4;14:300. doi: 10.1186/1471-2105-14-300. PMID: 24093637; PMCID: PMC3853019.
8. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. Bioinformatics. 2015 Jun 1;31(11):1738-44. doi: 10.1093/bioinformatics/btv060. Epub 2015 Jan 30. PMID: 25637562; PMCID: PMC4443681.
9. Malhis N, Wong ET, Nassar R, Gsponer J. Computational Identification of MoRFs in Protein Sequences Using Hierarchical Application of Bayes Rule. PLoS One. 2015 Oct 30;10(10):e0141603. doi: 10.1371/journal.pone.0141603. PMID: 26517836; PMCID: PMC4627796.
10. Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 2016 Mar;12(3):697-710. doi: 10.1039/c5mb00640f. PMID: 26651072.
11. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res. 2016 Jul 8;44(W1):W488-93. doi: 10.1093/nar/gkw409. Epub 2016 May 12. PMID: 27174932; PMCID: PMC4987941.
12. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018 Jun 1;34(11):1850-1858. doi: 10.1093/bioinformatics/bty032. PMID: 29360926.
13. He H, Zhao J, Sun G. Prediction of MoRFs in Protein Sequences with MLPs Based on Sequence Properties and Evolution Information. Entropy (Basel). 2019 Jun 27;21(7):635. doi: 10.3390/e21070635. PMID: 33267349; PMCID: PMC7515128.

14. He H, Zhou Y, Chi Y, He J. Prediction of MoRFs based on sequence properties and convolutional neural networks. BioData Min. 2021 Aug 14;14(1):39. doi: 10.1186/s13040-021-00275-6. PMID: 34391457; PMCID: PMC8364704.
15. Zhang F, Zhao B, Shi W, Li M, Kurgan L. DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. Brief Bioinform. 2022 Jan 17;23(1):bbab521. doi: 10.1093/bib/bbab521. PMID: 34905768.
16. Peng Z, Li Z, Meng Q, Zhao B, Kurgan L. CLIP: accurate prediction of disordered linear interacting peptides from protein sequences using co-evolutionary information. Brief Bioinform. 2023 Jan 19;24(1):bbac502. doi: 10.1093/bib/bbac502. PMID: 36458437.
17. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150-2. doi: 10.1093/bioinformatics/bts565. Epub 2012 Oct 11. PMID: 23060610; PMCID: PMC3516142.
18. Aspromonte MC, Nugnes MV, Quaglia F, Bouharoua A; DisProt Consortium; Tosatto SCE, Piovesan D. DisProt in 2024: improving function annotation of intrinsically disordered proteins. Nucleic Acids Res. 2024 Jan 5;52(D1): D434-D441. doi: 10.1093/nar/gkad928. PMID: 37904585; PMCID: PMC10767923.
19. Piovesan D, Del Conte A, Clementel D, Monzon AM, Bevilacqua M, Aspromonte MC, Iserte JA, Orti FE, Marino-Buslje C, Tosatto SCE. MobiDB: 10 years of intrinsically disordered proteins. Nucleic Acids Res. 2023 Jan 6;51(D1): D438-D444. doi: 10.1093/nar/gkac1065. PMID: 36416266; PMCID: PMC9825420.
20. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1): D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.