Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases

Seyed Ziaeddin Alborzi, David W. Ritchie, Marie-Dominique Devignes

Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases

Seyed Ziaeddin Alborzi[1,3], David W. Ritchie[1], Marie-Dominique Devignes[2,3]
[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine

INTRODUCTION

Protein domain structure classification systems such as CATH and SCOP provide a useful way to describe evolutionary structure-function relationships. Similarly, the Pfam sequence-based classification identifies sequence-function relationships. Nonetheless, there is no complete direct mapping from one classification to another. This means that functional annotations that have been assigned to one classification cannot always be assigned to another. Here, we present a novel content-based filtering approach called **CAPS (Computing direct Associations between annotations of Protein Sequences and Structures)** to systematically analyze multiple protein-domain relationships in the SIFTS and UniProt databases in order to infer direct mappings between CATH superfamilies, Pfam clans or families, and SCOP superfamilies. We then compare the result with existing mappings in Pfam, InterPro, and Genome3D.

MATERIALS & METHODOLOGY

Content-Based Filtering Intuition

User and Film Profiles

	Star1	Star2	Star3	Star4	Star5
User1	0	0	0	1	1
User2	1	1	0	0	1

	Star1	Star2	Star3	Star4	Star5
Film1	0	0	0	0	1
Film2	0	1	0	1	0

(Cosine) Similarity

	Film1	Film2
User1	0.7	0.5
User2	0.57	0.4

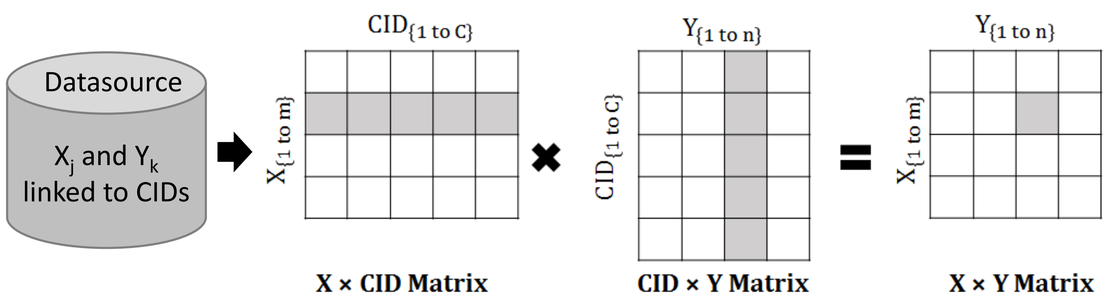
	CATH1	CATH2
SCOP1	0.7	0.5
SCOP2	0.57	0.4

SCOP and CATH Profiles

	CID1	CID2	CID3	CID4	CID5
SCOP1	0	0	0	1	1
SCOP2	1	1	0	0	1

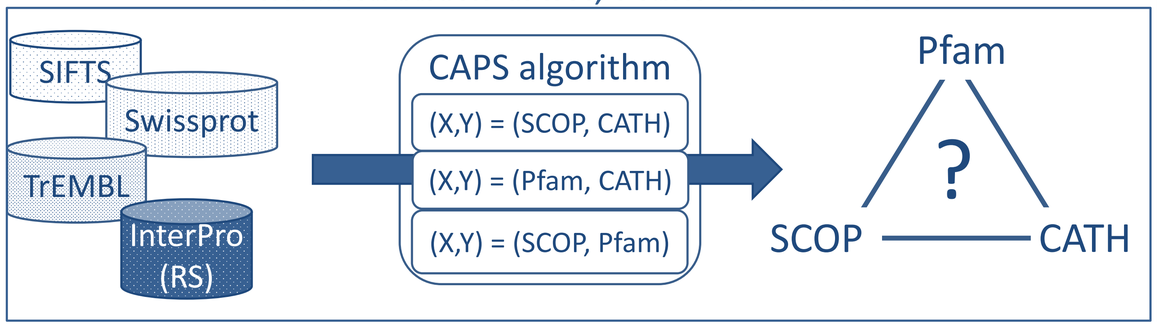
	CID1	CID2	CID3	CID4	CID5
CATH1	0	0	0	0	1
CATH2	0	1	0	1	0

Generalization: Scoring Matrix Calculation



X and Y : annotating entities; CID : Cluster Identifier for 100% identical sequences

Instantiation: CATH, Pfam and SCOP



The CAPS algorithm

Given

- X and Y, two sets of annotating entities,
- RS+, a reference set of confirmed associations between elements of X and Y,
- S_1 to S_n, n sources of relations between X and protein chains or sequences (CIDs), and between Y and protein chains or sequences,

1. For each datasource Si

- Extract X-CID and Y-CID relations (CID: clusters of 100% identical sequences)
- Compute dot product of normalized $(X \times CID)_i$ and $(CID \times Y)_i$ matrices to get the $(X \times Y)_i$ cosine similarity matrix for source S_i.

2. Aggregate similarity scores of all sources

$$ConfidenceScore_{X_j,Y_k} = \frac{\sum_i^n w_i S(X_j,Y_k)}{\sum_i^n w_i}$$

- Determine sources weights w_i that maximize AUC in ROC plots using RS positive examples of X-Y associations against background.

3. Determine the threshold confidence score

- Build a set of negative examples RS- (by random shuffling of relations supporting RS+) and build training and test sets of positive and negative examples
- Select confidence score threshold that maximizes F-measure on the training set of positive and negative X-Y associations and evaluate on the test set.

4. (Optional) Calculate P-Values for each association in each source S_i

- Hypergeometric law + Bonferroni correction
- Categorize the CAPS-inferred associations (Gold: all P-values significant ; Silver: more significant P-values than non significant ones ; Bronze: The rest).

Return

- "CAPS-inferred" X-Y associations (score > threshold), score and category.

CAPS-INFERRED ASSOCIATIONS

SCOP-CATH | Pfam-CATH | SCOP-Pfam

Table 1. CAPS mappings versus InterPro.

Dataset	SCOP-CATH Mappings	SCOP SupFam	CATH SupFam
Merged	580,763	1,851	2,604
CAPS	5,576	1,817	2,549
InterPro	2,856	1,637	2,231
Common with CAPS	2,764	1,634	2,225

Dataset	Pfam-CATH Mappings	Pfam Clans/Fam	CATH SupFam
Merged	1,068,601	7,228	2,754
CAPS	7,623	3,033	2,745
InterPro	3,573	2,008	2,494
Common with CAPS	3,494	1,998	2,489

Dataset	SCOP-Pfam Mappings	SCOP SupFam	Pfam Clans/Fam
Merged	1,004,741	2,111	7,165
CAPS	6,618	2,109	3,168
InterPro	2,100	1,537	1,752
Common with CAPS	2,053	1,532	1,745

Fig1. Distribution according to node degrees (Number of Associations)

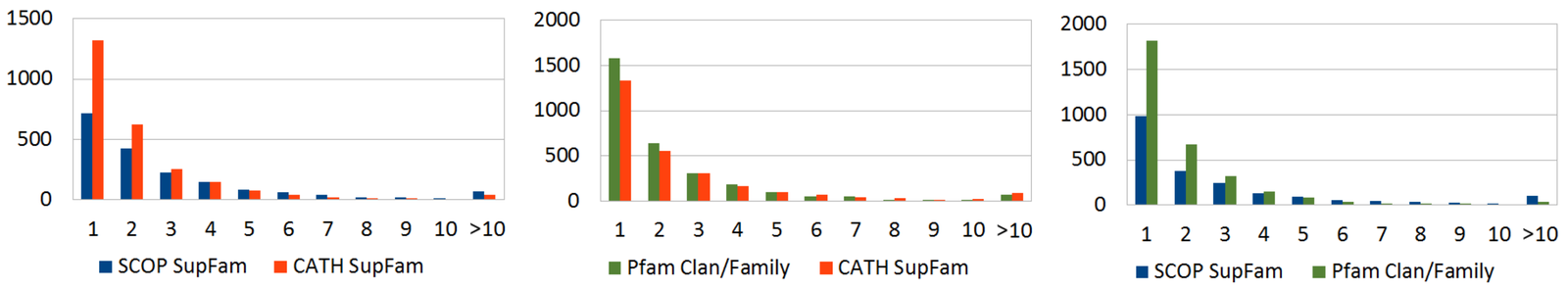
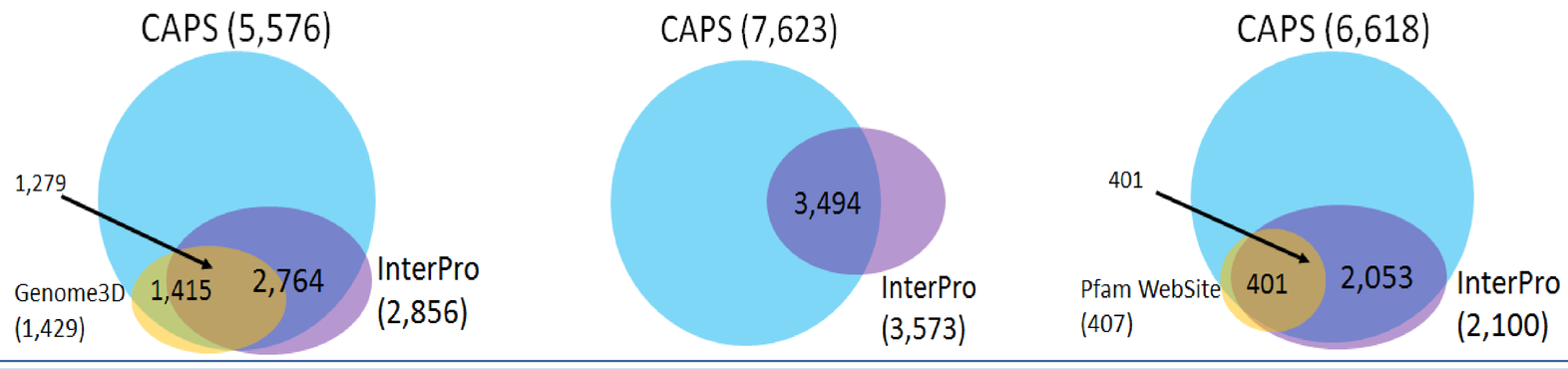
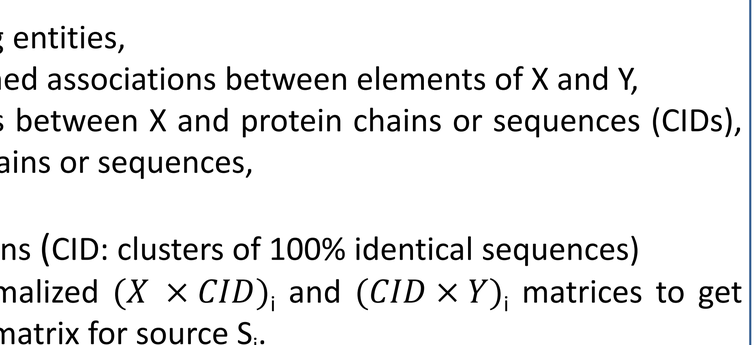


Fig 2. Intersection between our result (CAPS), InterPro, Genome3D, and Pfam website mappings.



TRIANGULAR VERIFICATION



	Mappings	SCOP	CATH
SCOP-CATH	5,576	1,817	2,549
Common with SCOP-{Pfam}-CATH	**5,438**	**1786**	**2518**
1:1 SCOP-CATH	506	506	506
Common with SCOP-{Pfam}-CATH	492	492	492

	Mappings	SCOP	CATH
Pfam-CATH	7,623	3,033	2,745
Common with Pfam-{SCOP}-CATH	**6,768**	**2629**	**2518**
1:1 Pfam-CATH	457	457	457
Common with Pfam-{SCOP}-CATH	393	393	393

	Mappings	SCOP	CATH
SCOP-Pfam	6,618	2,109	3,168
Common with SCOP-{CATH}-Pfam	**5,628**	1786	2629
1:1 Pfam-CATH	635	635	635
Common with Pfam-{SCOP}-CATH	478	478	478

CONCLUSION

- Over 90% of all associations found, are self-consistent with respect to triangular (SCOP-CATH-Pfam) associations.
- Overall, our approach finds 4 times as many SCOP-CATH superfamily associations than currently exist in Genome3D. These new associations will be beneficial to:
  1. Transfer annotations from one classification scheme to another.
  2. Investigate annotation consistency between different classifications.
- We are currently extending our approach to
  1. Analyze multiple associations in more detail.
  2. Confirm the associations using 3D structure alignment

LITERATURE CITED

1. F. C. Bernstein et al. The protein data bank. European Journal of Biochemistry, 80(2) :319–324, 1977.
2. R. D. Finn et al. Pfam: the protein families database. Nucleic Acids Research, 42(D1): D222–D230, 2014.
3. The GO consortium
4. A. G. Murzin et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology 247.4 (1995): 536-540
5. C. A. Orengo, et al. CATH–a hierarchic classification of protein domain structures. Structure 5.8 (1997): 1093-1109.
6. S. Velankar et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Research, 41(D1): D483–D489, 2013.
7. The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Research, 38(suppl 1): D142–D148, 2010
8. A. Mitchell et al. The InterPro protein families database : the classification resource after 15 years. Nucleic Acids Research, 43(D1): D213–D221, 2015.
9. T. E. Lewis, et al. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic acids research 41.D1 (2013): D499-D507.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Lorraine Region.

CONTACT

Zia Alborzi
INRIA Nancy Grand Est, LORIA, Bureau B133, France
Email: seyed-ziaeddin.alborzi@inria.fr
Website: www.loria.fr/~salborzi
Phone: +33 (0)7 83 29 89 83