Investigating the Link Between Intra-protein Interactions and Contiguous Hydrophobicity

By: Connor Pitman[1], Anthony Geneva[1,2] Matthew E. B. Hansen[4], Grace Brannigan[1,3]

[1]Center for Computational and Integrative Biology, Rutgers—Camden, NJ, 08102, [2]Department of Biology, [3]Department of Physics,4[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

Abstract

Though protein structure is the conventional link between protein sequence and function, structure is not directly accessible from the genome for all proteins. Predicting or interpreting the effects of mutations often requires identifying the local sequence context. Since protein structures are modular, the "local sequence" is frequently determined by the local secondary structure element. Some proteins, such as intrinsically disordered proteins (IDPs), can only be analyzed using their sequences. In cases like these, detecting innate modularity independent of secondary structure elements using only residue-level properties such as hydrophobicity and charge is extremely valuable. We have previously used contiguous hydrophobicity ("blobulation") to detect local sequence context for analysis of disease-associated mutations[1]. Here, we detail the blobulation algorithm and demonstrate its ability to detect subsequences associated with hydrophobic environments (core of a globular protein, transmembrane helices, etc.) We also present results from a coevolution analysis showing that these hydrophobic subsequences ("blobs") have evolutionary significance. We find that pairs of similarly hydrophobic blobs are enriched for coevolving residues. Additionally, we find that within these blobs, the types of coevolving amino acid pairs change depending on the blob containing them, suggesting that blobulation provides a meaningful framework for defining the context around coevolving pairs and could be useful in further coevolution studies.

Blobulation

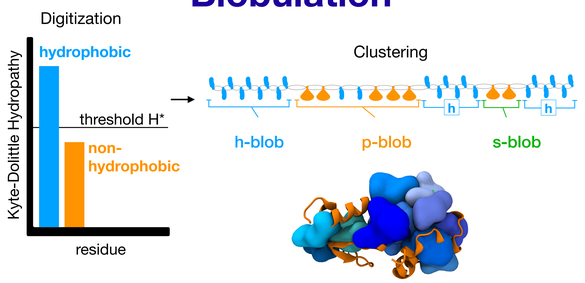
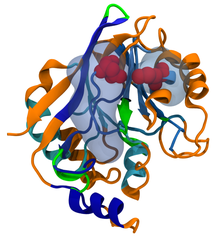


Figure 1: Blobulation, our algorithm for detecting intrinsic modularity in protein sequences based on hydrophobicity. The algorithm involves two steps: digitization using hydrophobicity threshold H* (left), and clustering (middle). Figure adapted from [1]. Example representation made in VMD of Lysozyme blobs (right, Uniprot: P00720, PDB:2LZM).

Detecting interactions from sequence

- Pairs of residues found at coevolving sites (two positions in orthologous proteins consistently co-ocurring across evolutionary history) are often found in contact [2]
- Additionally, residue-level properties such as hydrophobicity and the charge class of residue groups provide information about a protein's ensemble [3, 4]
- Highly charged regions either attract or repel each other, while neutral regions tend to be globular [3]

Figure 2: Representation of coevolving residues in a globular protein. The protein backbone is colored by blob, with the coevolving residue pair in red and the blobs containing this pair are transparent.



Research Questions

1. Are coevolving residue positions likely to be located within hydrophobic hot-spots (h-blobs)?
2. Is the viability of a hydrophobic-to-hydrophobic mutation dependent on…
   - the identity of other hydrophobic residues?
   - the presence of surrounding hydrophobic residues?

Approach

1. Detected coevolving sites in a large Bacterial protein dataset (1630 protein families, with ~229 orthologues per family - previously used to investigate the role of structure in coevolution) using CoMap [2]. CoMap detects coevolving pairs of sites among orthologues using a clustering approach.
2. Blobulated protein sequences from above (as in Figure 1).
3. Calculated enrichment (equation below) of coevolving residues for all blob type pairings, and for all amino acid type pairings among varying blob types ("contexts"). Null expectation was generated using a permutation test.

N_{ab}^{obs} = Number of detected coevolving pairs 'ab'

N_{ab}^{perm} = Null frequency of pair 'ab', generated by random shuffling of sites (as in Approach)

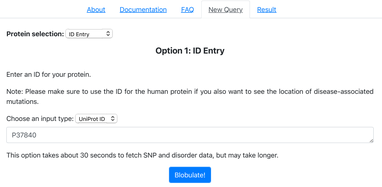
$$\text{Enrichment} = \frac{N_{ab}^{\text{obs}}}{N_{ab}^{\text{perm}}}$$

Summary

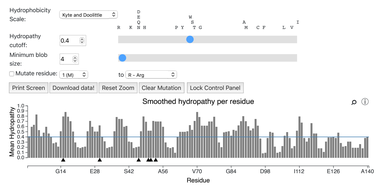
1. Pairs of similarly hydrophobic blobs (h-h and p-p), are enriched for coevolving residues.
2. Pairs found in the same context (h-h and p-p) tend to be enriched for charged residue pairs and some highly hydrophobic residues than those in opposite contexts (h-p), but there are notable exceptions (C-C, C-Y, and some aromatic pairs).
3. The amino acid composition among various blob pairs is distinct, though some residue pairings (ex. G-G and E-K) are enriched across all blob pairs.

Blobulate your own Proteins - Web Interface

blobulator™
hydrophobic modularity in protein sequences
Brannigan Lab • Rutgers University – Camden



1. Protein input options
   - Uniprot ID
   - Ensembl ID
   - amino acid sequence



2. Set the hydrophobicity scale, hydropathy cutoff, and minimum blob length

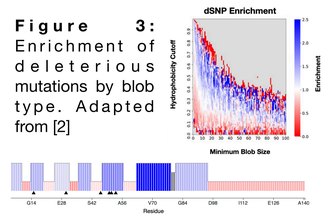


Figure 3: Enrichment of deleterious mutations by blob type. Adapted from [2]

3. Browse the blobulated results including enrichment of deleterious mutations

Blobulator.brannigan-lab.org

Blobulate your own protein here!

doi:10.1101/2024.01.15.575761

Are hydrophobic blobs enriched for intra-protein coevolving sites?

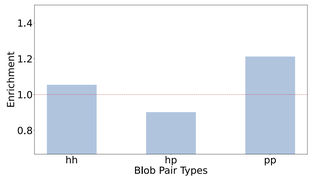


Figure 4: Enrichment of each h- and p- blob-type pairing for coevolving residues. All pairs are significant at an FDR of 0.05. Horizontal red line represents boundary between enrichment and depletion (1.0). Blobulation was done using a hydrophobicity threshold of 0.4 and a minimum length requirement of 4.

Does the amino acid composition of coevolving sites differ by blob type?

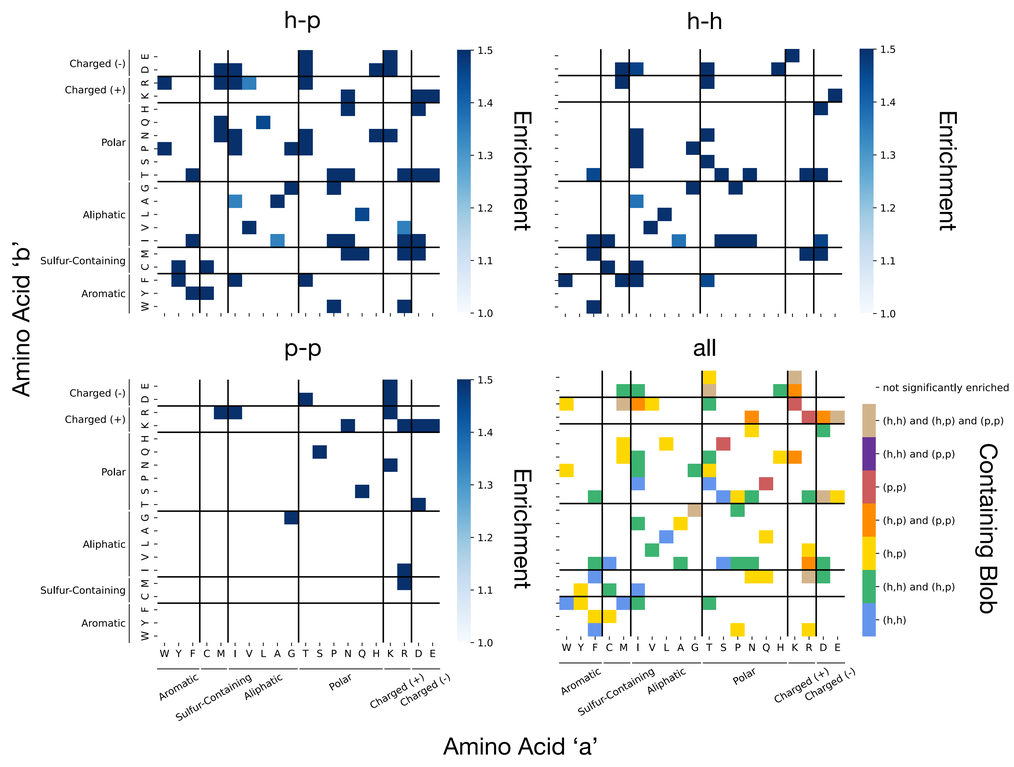


Figure 5: Amino acid pairs categorized by the type of blob that contains the same sequence position in the ancestral state: h-p (top left), h-h (top right), p-p (bottom left), and whether each filled bin across plots is unique to or present in each (bottom right). Blobulation was done using the same settings as Figure 4. Black lines delineate between aromatic, sulfur containing, aliphatic, polar, and charged residues.

Acknowledgements

•Rutgers Office of Advanced Research Computing (OARC)
•NRT, NSF DGE 2152059
•NIH 1R35GM134957

References

[1] R. Lohia, M. Hansen, G. Brannigan. PNAS, 2022.
[2] S. Chaurasia and J. Dutheil. Molecular Biology and Evolution, 2022.
[3] R. Das and R. Pappu. PNAS, 2013.
[4] V. Uversky, J. Gillespie, A. Fink. Proteins, 2000.