Untangling the role of local sequence context and identity in residue coevolution

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

Protein evolution is constrained by intra-protein interactions. Detecting coevolving positions in proteins allow us to elucidate these interactions and better quantify the underlying mechanisms that drive them. In proteins, many structural and functional features are stabilized via groups of neighboring residues interacting as a unit, often represented by secondary structure elements. However, coevolving residues tend to fall outside of secondary structure elements. Here, we use a method developed by our lab – blobulation – to define the local sequence context ("blobs") around coevolving pairs via contiguous hydrophobicity. We investigate the role of this context by identifying the types of blobs that coevolving pairs are likely to be found in and whether residues with certain biochemical properties and certain amino acids are more likely to members of coevolving pairs than random chance, given their local context. We find that 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.

Research Questions

Main question: Can we use blobs in place of secondary structure to define the local sequence context around coevolving residues?

Sub-Questions

1. Do some blob types contain coevolving residues more often than would be expected by random chance? And if so, which ones?

2. Do the types of coevolving amino acids vary depending on the blob containing them, or are they the same regardless of containing blob?

Coevolving Positions

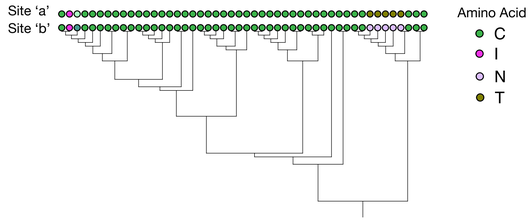

Figure 1: Coevolving positions. Pairs of positions that co-vary across evolutionary time are considered coevolving. The amino acid identities of two coevolving sites across an example set of species are shown.

- Coevolving residues (shown in figure 1) are often found in contact and outside of secondary structure elements [1].
- Properties such as hydrophobicity and the charge class of residue groups provide information about a protein's ensemble [2, 3]
- Coevolving residues studies have traditionally focused on oppositely charged residue pairs, without incorporating the role of the local sequence context.

Blobulation

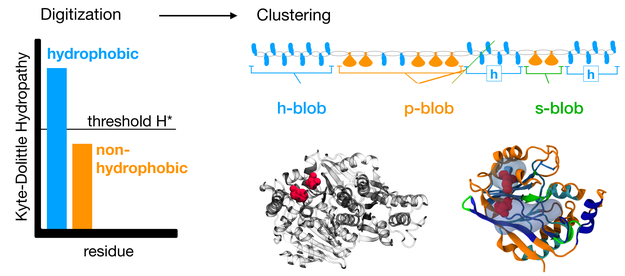

Figure 2: Blobulation, our algorithm for detecting intrinsic modularity in protein sequences based on hydrophobicity. The algorithm involves two steps: digitization using hydrophobicity threshold H* (left), and clustering (middle). Figure adapted from [4]. Example of coevolving residues (red) in an unblobulated protein (left) and a blobulated protein (right).

What types of amino acids tend to coevolve?

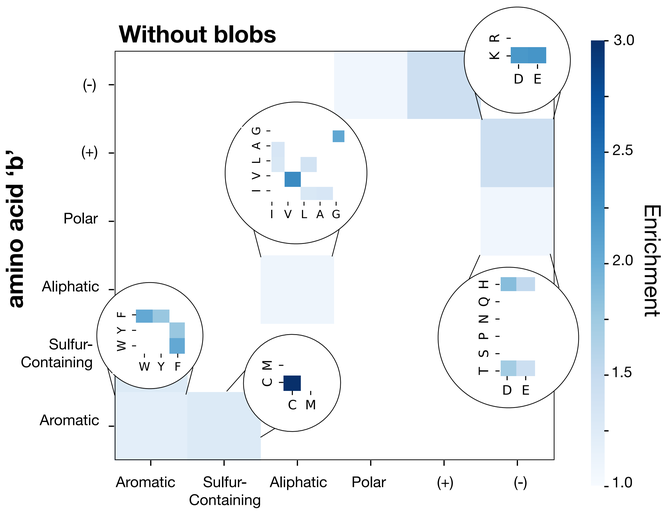

Figure 3: Enrichment of pairs of biochemical properties (left), and amino acid pairs (as pop-outs), for coevolving positions. Significantly enriched (FDR > 0.05) residues and types are shown in blue.

How do interactions between coevolving amino acids depend on their local sequence context?

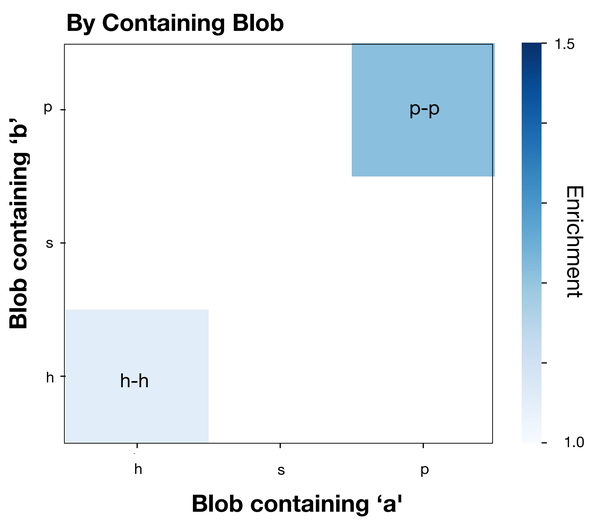

Figure 4: Enrichment of pairs of blob types (as defined in Figure 2) for coevolving positions. Significantly enriched (FDR > 0.05) residues and types are shown in blue.

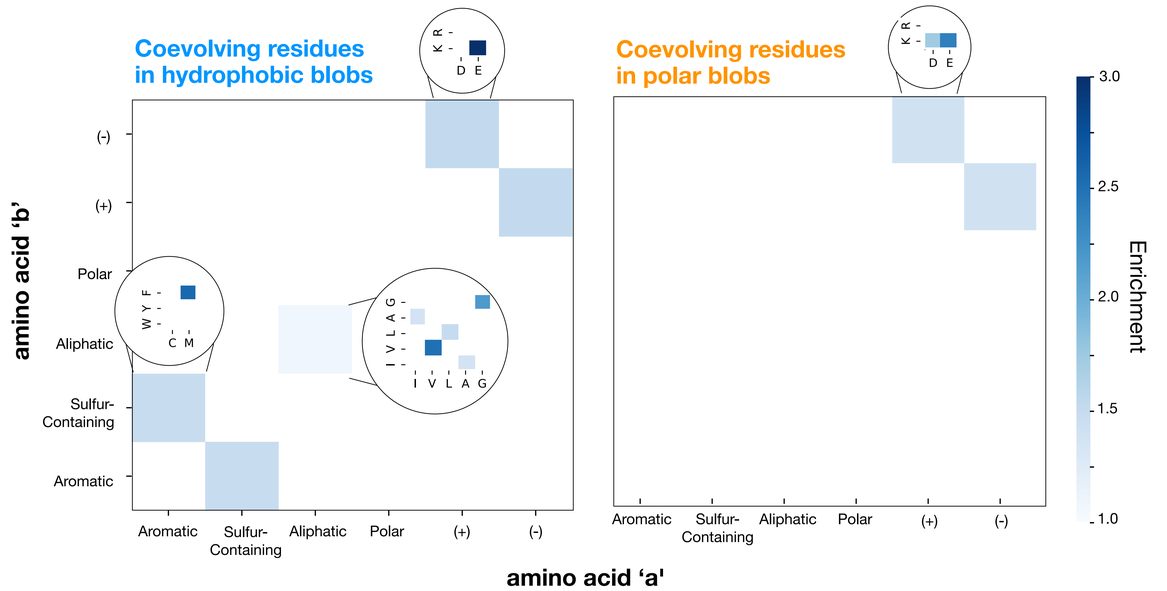

Figure 5: Enrichment of biochemical properties and amino acid pairs (as pop-outs) both in h-blobs (left) and both in p-blobs (right) for coevolving positions. Significantly enriched (FDR > 0.05) residues and types are shown in blue.

Approach

1. Detected coevolving sites in a large Bacterial protein dataset (1630 protein families, with ~229 orthologs per family - previously used to investigate the role of structure in coevolution) using CoMap [1]

2. Blobulated all protein sequences (as in Figure 1)

3. Calculated enrichment 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. All detected coevolving sites were shuffled, with the amino acids found at each remaining unshuffled.

Enrichment

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

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

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

Summary

1. Coevolving pairs are more likely than by random chance to both either be found in h-blobs or in p-blobs, indicating both that:
   - In coevolution studies, where secondary structure cannot be used to define "local sequence context", blobs can.
   - h- and p- blobs represent modules that interact with other blobs of the same type.
2. h-blobs contain interactions under evolutionary constraint not present in p-blobs, between:
   - aromatic and sulfur-containing residues
   - aliphatic residues
3. Oppositely charged interactions are constrained regardless of blob type

Acknowledgements

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

Blobulator.brannianlab.org

Blobulate your own protein here!

References

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