

---

# Exploring Structural Implications of Positional Dependencies in Protein Sequence Alignments

Hatice Gulcin Ozer\*

Biophysics Graduate Program, The Ohio State University

William C. Ray

Children's Research Institute and The Department of Pediatrics, The Ohio State University

---

## Introduction

Predicting physical distances between amino acids in protein alignments provides invaluable information towards anticipation of their complete 3-dimensional structure. Extracting constraints using only sequence information is an indispensable direction, since the number of known protein or nucleic acid sequences grows much faster than the number of known 3-dimensional structures. Detecting interpositional dependencies within the multiple sequence alignments of protein families and understanding their physical consequences will be a big step in this direction.

In studying positional dependencies, we observed that dependencies are often the result of physical proximity. Since physicochemical interactions between many identities in the biomolecule are involved in proper folding and functioning, it is expected to observe dependencies amongst some positions. Therefore, identification of statistically significant interpositional dependencies within family alignment will further assist researchers to determine constraints on family structure.

In this study, we examined the critical parameters of interpositional dependencies, also called pairwise correlations, to estimate structurally important residues for family alignments.

## Methods

Interpositional relationships for a given family alignment are estimated via subtracting the expected number of occurrences from observed occurrences for every possible pair of positions and identities in the alignment. This measure is the statistical residual.

In this study, we compared the magnitudes of these residuals and their statistical significances to actual physical distances of corresponding residues. First, we generated all possible pairwise correlations for all family alignments in Pfam (Protein and domain families database of alignments). Then, we examined the actual physical distances between correlating positions using the known PDB structures for the family.

for acceptable statistical

Pfam (*ver 21*) includes 8,957 family alignments and 2,588 of them refer to at least one PDB structure. To be able to statistically sound, we computed pairwise dependencies and their statistical significances for only 1,746 Pfam families and their corresponding physical distances on 12,812 referred PDB structures. We also calculated ratio of actual physical distances to one-half of the molecule's average diameter. These ratios allow us to investigate whether the correlating residues are

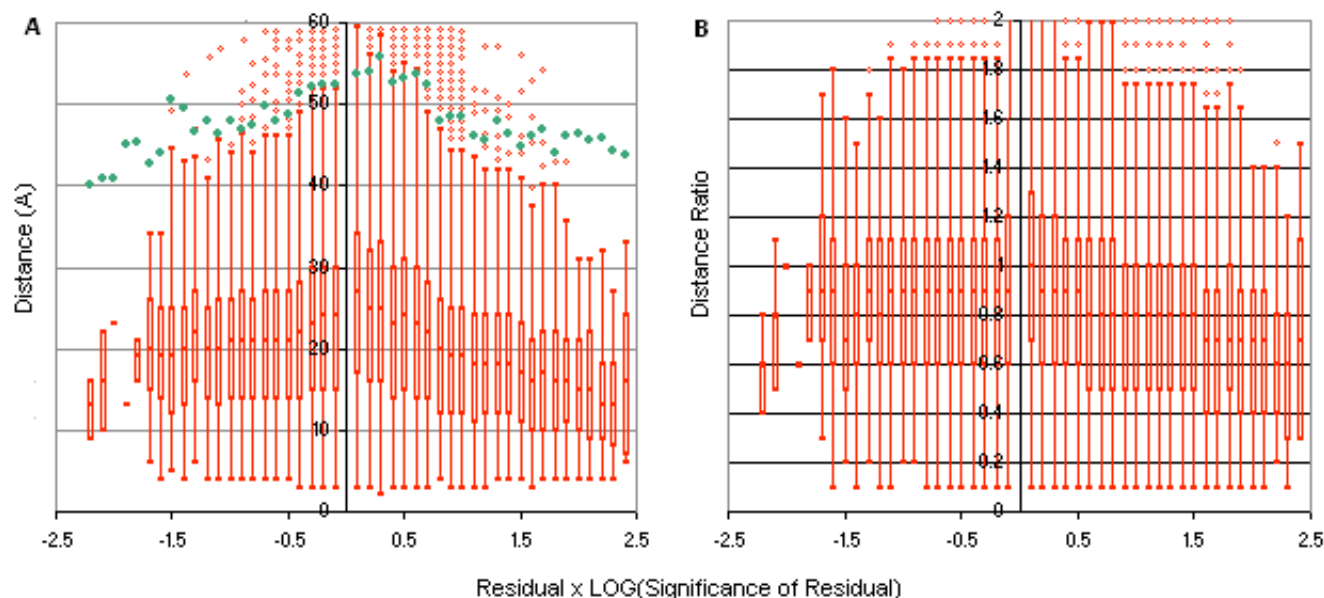
at least in the same 3-dimensional quartile of the molecule.

## Results

We observed that actual physical distances between correlating pairs of residues decrease as the corresponding residuals and their significances increase. This relationship gets stronger if we use ratios of physical distances to one-half of the protein's average diameter.

We combined two main parameters, i.e. residual and residual's statistical significance, by multiplying the residual and the logarithm of residual's statistical significance. Thus, large values of the combined parameter imply large and statistically significant residuals. Figure 1 presents the box plots of physical distances and distance ratios of correlating residues over this combined parameter. For the large values of this combined parameter, we observed that about 75% of the actual physical distances of the correlating residues are smaller than 20Å, and more than 75% of the distance ratios of the correlating residues are smaller than 1.

These findings allow us to argue that large and significant residuals strongly suggest physically being in the same 3-dimensional quartile of the molecule.



**Figure 1. Box plots of physical distances (A) and distance ratios (B) of correlating residues.** (A) depicts box plot of the actual physical distances between correlating residues over the combined parameter [Residual x LOG(Significance of Residual)]. Green circles represents the average diameters of the molecules. (B) depicts box plot of ratio of the distances between correlating residues to the half of the molecule's average diameter over the combined parameter. In both (A) and (B), there are a few number of outliers beyond the depicted scale for the small values of the combined parameter.

## Conclusion

We observed a significant relation between positional dependencies and physical distances between these positions in the molecule's structure. This allows researcher to predict physical distances of identities found to be dependent in multiple sequence alignment of proteins families via computational

analysis. The acquisition of such knowledge is critical to model overall 3-dimensional structure of the proteins.

---

\*Corresponding author: ozer.9@osu.edu