

Protdist: An analysis of error...

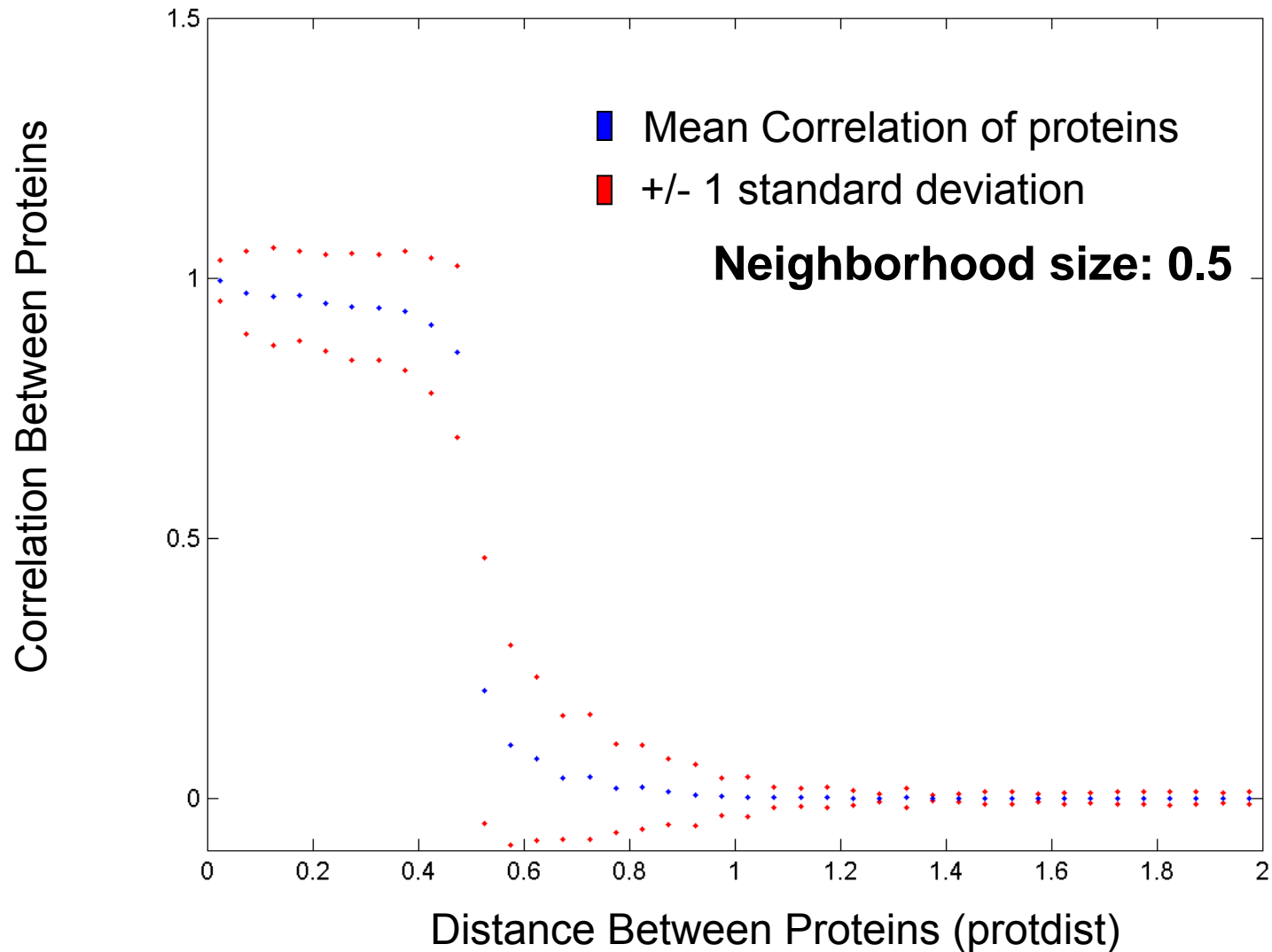
And the implications

Method

- Using a sample of all 400 phage from the 2004 GenBank® database, sequence DNA and generate all possible proteins that could be coded for.
- Using Protdist, calculate the distances between each pair of proteins.
- Protdist – A distance measure that measure distance in terms of the “average number of changes per amino acid” between 2 proteins.
- Run statistical analyses on data and look for possible implications.

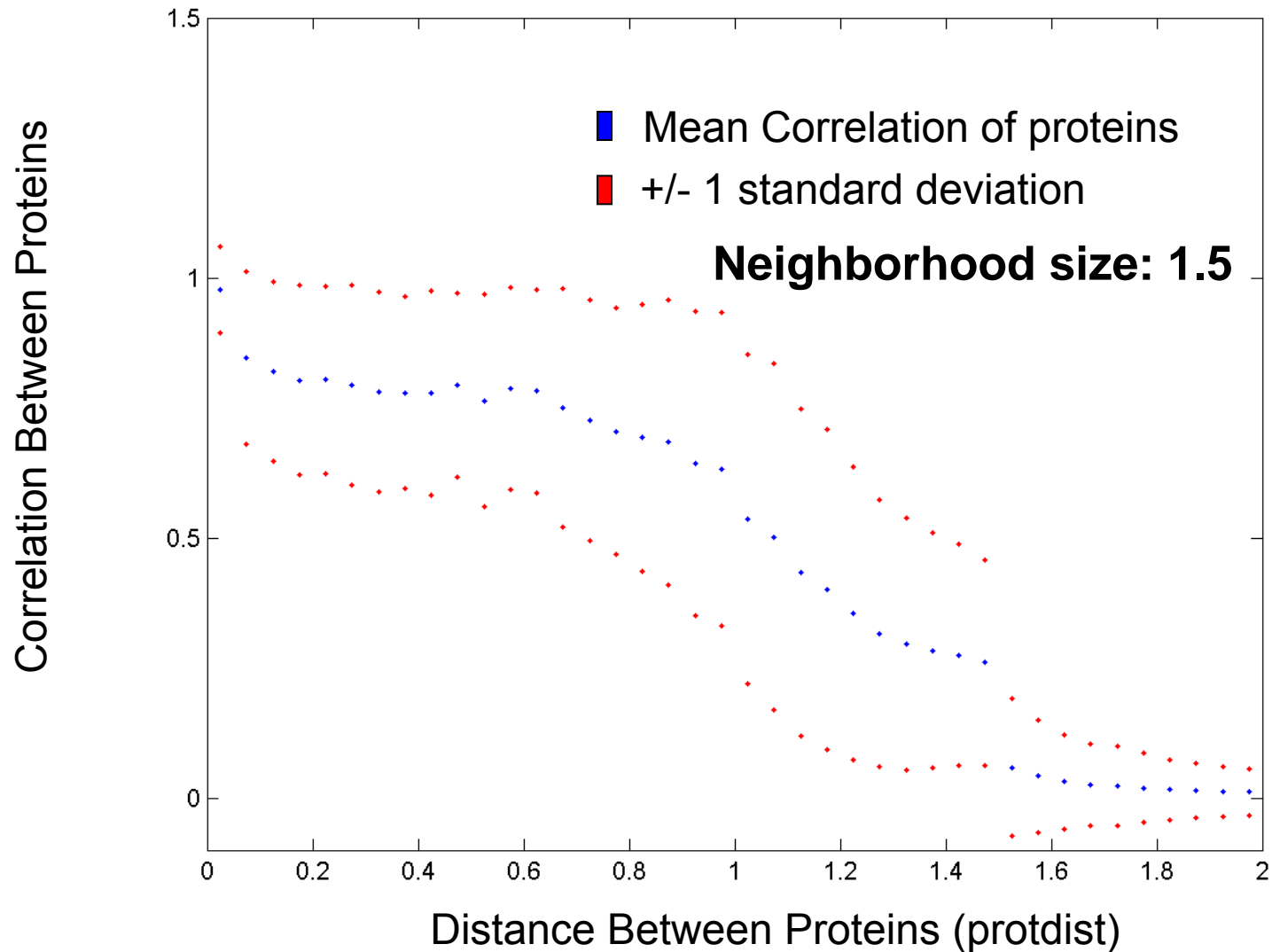
Results and Implications (Part I)

-Treelike behavior at small distances-



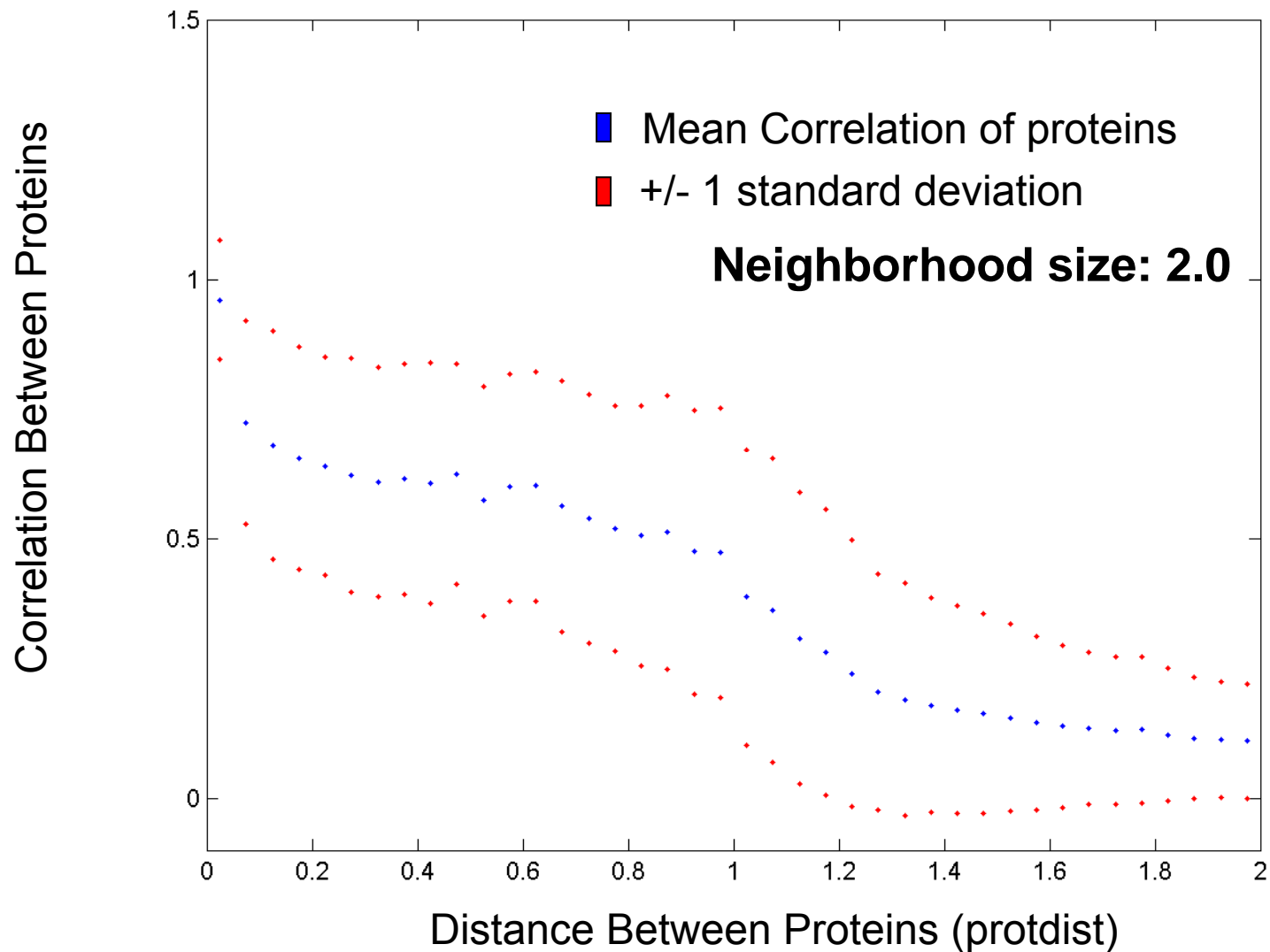
Results and Implications (Part I)

-Treelike behavior... begins to disappear after 1.5



Results and Implications (Part I)

-Treelike behavior is gone

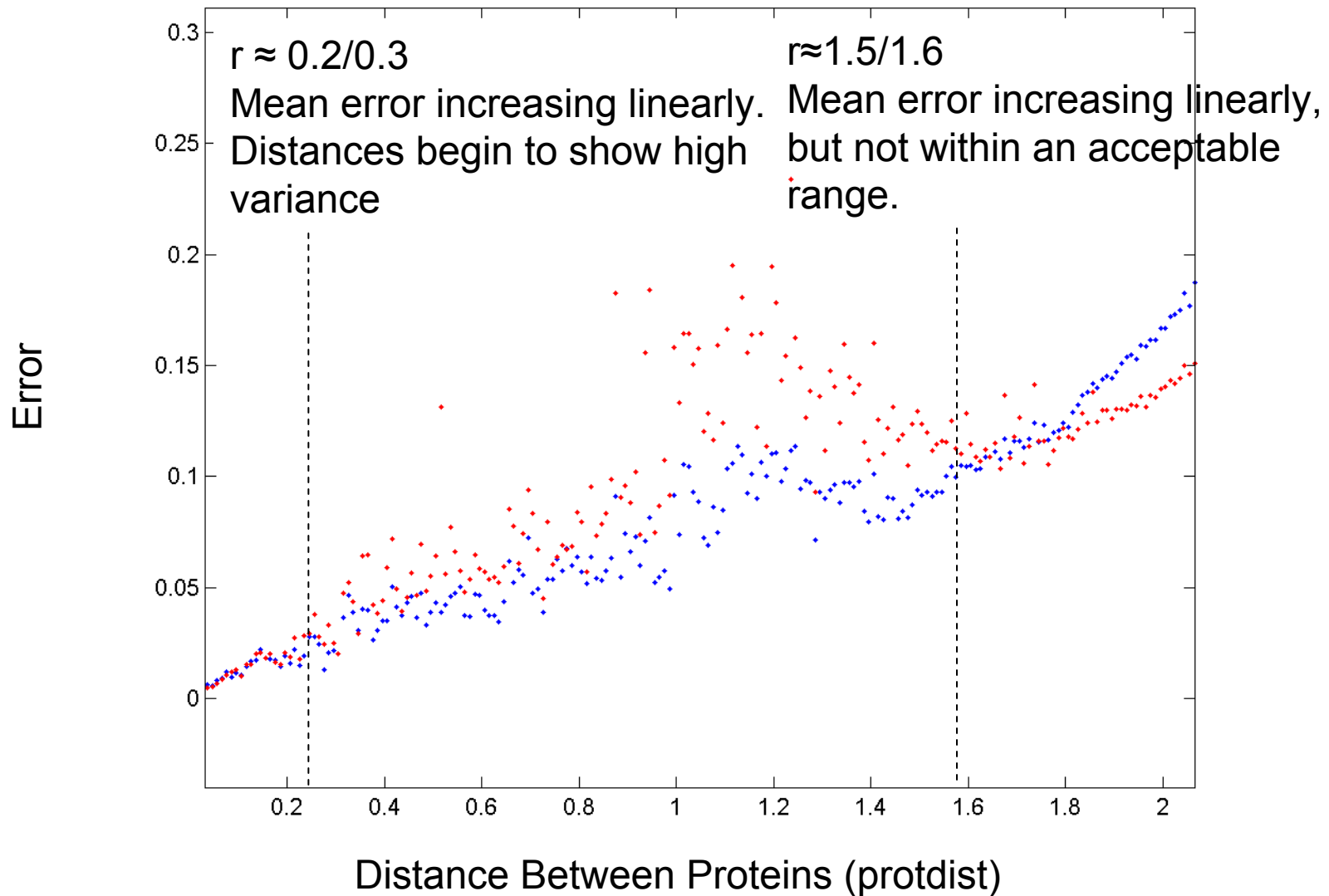


Another Statistical Analysis

- Previously, we inferred that treelike structure was preserved for neighborhoods of sizes less than about 1.5
- Alternatively, assume treelike/ultrametric structure is always present, and determine adherence to this assumption depending on the distance between the proteins.

Results and Implications (Part II)

Defines Distinct Regions

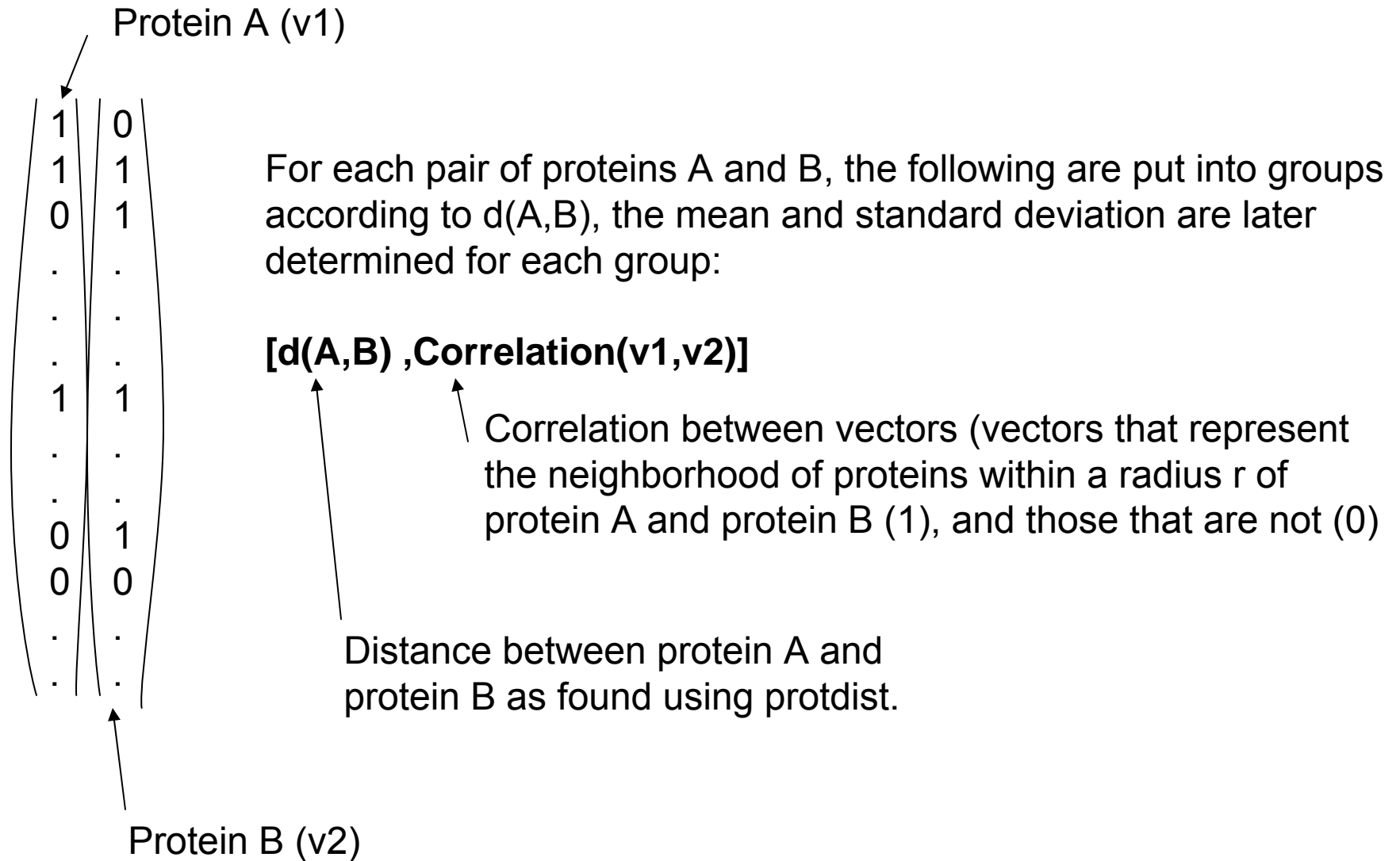


Conclusion

- The relationship between the proteins that are found in a subsample of phage is treelike at small distances (<1.5).
- In so much as protdist was used to measure distances between pairs of proteins, the error in terms of deviation from a treelike structure becomes unacceptable around 1.5 or 1.6

The End

Method of Finding Correlation



Radius= 0.1

● Mean Correlation
● +/- One Standard Deviation

