Clustering of Multidimensional Data Sets with Applications to Spatial Distributions of Ribosomal Proteins

REU Site: Interdisciplinary Program in High Performance Computing

Nil Mistry[1], Jordan Ramsey[2], Benjamin Wiley[3], and Jackie Yanchuck[4],
Graduate RAs: Xuan Huang[5] and Andrew Raim[5],
Faculty Mentors: Matthias K. Gobbert[5] and Nagaraj K. Neerchal[5],
Client: Philip J. Farabaugh[6]

[1]Department of Mathematics and Statistics, University of Connecticut
[2]Department of Computer Science and Electrical Engineering, UMBC
[3]Department of Mathematics and Statistics, University of New Mexico
[4]Department of Mathematics, Seton Hill University
[5]Department of Mathematics and Statistics, UMBC
[6]Department of Biological Sciences, UMBC

Technical Report HPCF–2013–10, www.umbc.edu/hpcf > Publications

Abstract

Consider ribosomal proteins, each with a three-dimensional spatial location. Proteins related to the cofactor phenotype may be randomly or non-randomly distributed within the ribosome. To investigate this question, the Mahalanobis distance is computed between each pair of protein locations, and the optimal pairing is determined by minimizing the sum of the within-pair distances. Since no single code exists that allows for the computation of Mahalanobis distances, determining the optimal pairing, and determining whether the two groups are statistically different, we created a code that allows a user to do just this. The user can also compute an exact p-value for this distribution rather than rely on an approximation.

1 Introduction

Multivariate distributions are analyzed under different conditions than one-dimensional distributions which may be considered under the Wilcoxon Mann-Whitney test. Several methods exist to analyze such sets; in particular, [3] has introduced an exact, distribution-free test which compares multivariate distributions based on adjacency. Conveniently, one need not worry about the normality of a data set to rigidly uphold other standard methods for comparing distributions, and therefore this method offers a convenient algorithm for producing a conclusion regarding adjacency of multivariate data without having to worry about such details.

Our client Dr. Philip Farabaugh provided data on protein locations. His experimental data is a set of three dimensional coordinates for proteins contained within a ribosome. Each protein is either phenotype-related or not.

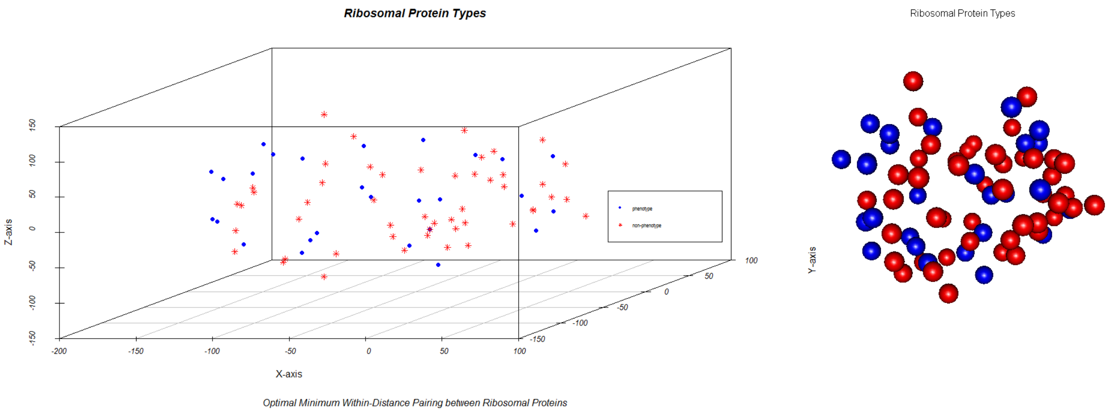

Figure 1.1: Ribosomal protein types.

In our case, two sets of ribosomal proteins exist, one containing phenotype traits and another containing non-phenotype traits, whose distributions concerning adjacency between groups may be interesting to consider. More specifically, the positions of each protein in a three-dimensional environment compare the clusters of each set, and analysis of relations between these two groups yield conclusions regarding the similarity of the two sets, as seen in Figure 1.1. For instance, dissimilar distributions between categories could indicate dissimilar traits between phenotype and non-phenotype categories.

Comparing the positional distributions to one another in two dimensions is possible if one considers only two axes at a time. Figure 1.2 shows a scatter plot, histogram, and box plot for each comparison between any two axes. The p represents phenotype traits present within the proteins, and correspond to the blue color, while the n represents the non-phenotype proteins in the pink color. One may notice that each comparison appears roughly similar in distribution.

Since variance within groups may have an effect on distance, Euclidean distance is not necessarily an appropriate measurement of adjacency. That is, if one set has a high variability among elements, or protein types, in comparison to the second group, then a large distance in the first may be an exaggeration to the second. To account for this possible discrepancy, Mahalanobis distance is utilized, which computes the Euclidean distance with respect to the variability of the two sets. In effect, this is obtained by computing the covariance matrix between the two groups, and results in an accurate description of adjacency concerning the sets.

Once the Mahalanobis distance is computed between each possible ribosomal protein pair, the optimal, minimum sum of within-distances is requested in an attempt to obtain the most accurate pairing for each couple. As a result, the adjacency of the phenotype set to the non-phenotype set may be fully considered as members are compared on a large scale to perceive differences in position. Indeed, if many non-matching pairs are determined as closest pairs through this process, then there is a higher expectation for similarity between phenotype and non-phenotype positional adjacency. To compute these closest pairs, an

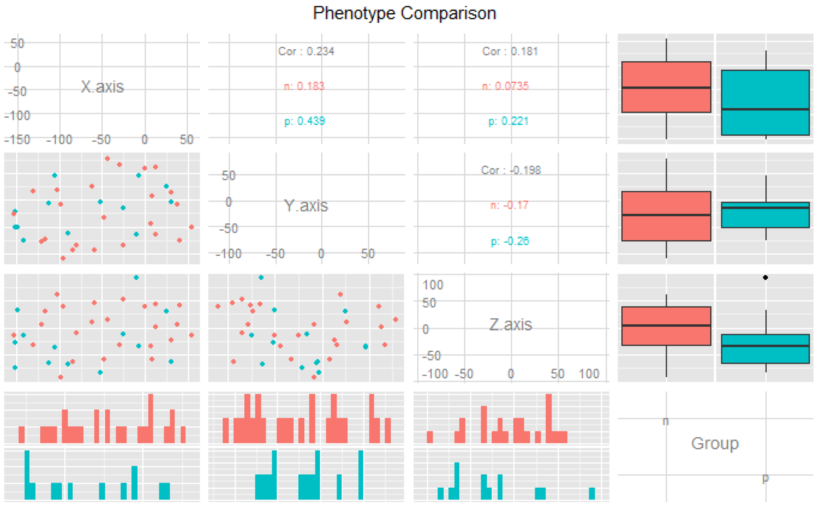

Figure 1.2: Two-dimensional comparison.

optimal, non-bipartite matching algorithm is used to sort through each pairing design and determine the shortest distance between groups.

To yield conclusive evidence regarding the adjacency of phenotype and non-phenotype groups, a p-value may be determined based on the number of non-matching pairs explained in the previous paragraph. Intuitively, a greater number of non-matching pairs should provide evidence to similar adjacency between groups, whereas a low number of non-matching pairs should likewise lead one to expect dissimilar adjacency. Deriving the exact null distribution to the set of non-matching pairs, or the probability density function with random variable A_1 referring to these non-matched pairs, the null hypothesis upholding equal adjacency between groups may be rejected if A_1 is small enough [3]. By this, a z-score may be obtained, and consequently a p-value, to easily offer a conclusion regarding the adjacency of groups.

2 Background

Univariate data may be completely ordered by scaling elements with respect to a single variable, allowing for straightforward comparisons within the sample. Further, two univariate groups may be easily compared against one another using the Wilcoxon signed-rank test so long as certain assumptions have been met. For instance, pairs being compared must be chosen randomly and independently, and the distribution of data from either group must be symmetric about the median; however, neither distribution must be normally distributed, thereby admitting a wider range of data to be analyzable under this method [1]. Consequently, the Wilcoxon signed-rank test is referred to as a non-parametric test under the above permission.

Nevertheless, multivariate data provides a less straightforward example of group comparison. For example, while a single variable from two different sets, such as length, may be relatively simple to order on a scale, one may have more difficulty comparing multiple variables, such as age and height, from two different sets. That is, while 6 cm is longer than 5 cm, how may one compare 6 cm north and 4 cm west with respect to a center point, in relation to 5 cm south and 3 cm east? Indeed, such comparisons may not be completely ordered. Comparing entire data sets with respect to each other might provide a better approach, but these "neighbor count" based statistical methods are not distribution free.

Friedman and Rafsky (1979) implemented pairwise distances to construct a minimum spanning tree, and removed edges in the tree that connected the two different groups. Accordingly, the resulting number of remaining disjoint sub-trees may be utilized as a test statistic. Another test created by Schilling (1986) and Henze (1988) paired up data using the nearest neighbor to each subject, and then counted the number of times that subjects from the same group were paired together. However, these tests are not distribution-free [3]. To be distribution free, the distribution of the test statistic should be a known distribution that depends on the sample size, but not on the distributions of the input data.

In 2005, Rosenbaum developed a test statistic to analyze multi-dimensional data that was distribution free. Computing the Mahalanobis distance within groups, thereby taking into account variance of each variable, one may optimally divide N data points into $\frac{N}{2}$. Each pairing must contain either zero, one, or two points from the first group, and a test statistic referring to the equality of distributions is produced from the number of pairs with exactly one data point from the first group.

The ideal pairing is produced from optimal, non-bipartite matching. That is, the sum of the within-pair distances must be minimized, and to accomplish this, each distance must be considered in relation to the others. The number of arithmetic operations required to derive the Rosenbaum test statistic is $O(N^3)$ [3]. For example, Papadimitriou and Steiglitz (1982) provides several various algorithms, such as the Hungarian Method and the Weighted Matching Algorithm. In this paper, implementation of optimal, non-bipartite matching is obtained from E. Rothberg's C algorithm [4].

Applying the above process to a biological application, this paper considers the relation of distributions between two groups of RNA proteins. About 80 proteins exist within the structure of the ribosome, each assuming its own unique position. Certain proteins are carriers of a particular phenotype, while others are not and are identified in this study as non-phenotype proteins. Phenotypes are physical manifestations of a particular characteristic that results from a particular genotype and its relationship with the surrounding environment. An important question to ask whether the proteins with and without the specific cofactor phenotype are randomly distributed throughout the RNA molecule or not. This question is answered for a specific data set containing two groups of RNA proteins, phenotype and non-phenotype, and the similarity of their positional distributions is considered.

3 Numerical Methods

3.1 Mahalanobis Distance

The Mahalanobis distance computes a non-Euclidean distance between pairs that are located within groups of distinct data types, with respect to the pooled variance-covariance matrix of the groups. Let p characteristics in $X = (X_1, X_2, \ldots, X_p)^T$ be discriminated among k groups. Then, let $\bar{X}_i = (\bar{X}_{i1}, \bar{X}_{i2}, \cdots, \bar{X}_{ip})^T$ be defined as the vector of mean responses for the i^{th} sample. Further, let S_i be defined as the p-by-p variance-covariance matrix for the i^{th} sample. Then, the pooled variance-covariance matrix is defined as

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_k - 1)S_k}{n - k} \tag{3.1}$$

The Mahalanobis generalized squared distance (referred to hereon as Mahalanobis distance) from a particular observation X to the center of the i^{th} sample is

$$D_i^2 = (X - \bar{X}_i)^T S^{-1} (X - \bar{X}_i), \tag{3.2}$$

where it is understood that $(X - \bar{X}_i)^T$ is the transpose of $(X - \bar{X}_i)$, and S^{-1} is defined as the matrix inverse of S [1].

One may notice that redefining the covariance-variance matrix as the identity matrix, the above equation reduces to the 2-dimensional Euclidean norm, referring to Euclidean space. In this report, the Euclidean norm is also determined to consider the presence of variance within the phenotype and non-phenotype groups, and this result is compared to the Mahalanobis conclusion stated above. Consequently, one may imagine the Mahalanobis distance as similar to the Euclidean, only also taking into account variance within the set.

In our case, we consider two groups of RNA proteins specifically distinguished by some elements containing phenotype traits and others containing non-phenotype traits. In this instance there are two groups, and therefore the Mahalanobis distance must take into account the distance between the centers of the phenotype and non-phenotype clusters. Accordingly, the Mahalanobis distance from the i^{th} group to the j^{th} group reflects the Mahalanobis distance between the centers of each group, and is defined as

$$D^2(i,j) = D^2(j,i) = (\bar{X}_i - \bar{X}_j)^T S^{-1} (\bar{X}_i - \bar{X}_j). \tag{3.3}$$

This method allows us to derive accurate distances between closest pairs within the entire dataset, taking into account variance within each separate group.

3.2 Distance Computation

R code was developed that takes the length of the data and ranks each vector of data being analyzed. In this case, ranks were computed for the x-, y-, and z-coordinates. In order to break ties in ranks, we took the minimum of all ranks included in the tie. For example, if there was a tie for first, second, and third place, then all elements included in the tie would receive a rank equal to one. The covariance matrix, S, is computed for these ranks. Since we know the ranks and covariance matrix, we use the Mahalanobis distance for every combination between pairs within the matrix. The computed data is output to a text file in order to read it into C code.

3.3 Non-Bipartite Matching

The C code is an optimal non-bipartite combinatorial optimization matching sorting algorithm. The optimal pairing is determined by minimizing the sum of the within-pair distances between ribosomal proteins. In this case, there are $\binom{N}{2}$ total possible pairings, where N is the total number of proteins, in this case, 76. The minimum distance that is calculated divides 76 proteins into 38 non overlapping pairs, and minimizes the sum of the 38 distances within the 38 pairs. From this data we develop R code to match protein pairs, based on phenotype and non-phenotype matches. Based on the proportion of phenotype and non-phenotype proteins within the data set using the number of non-matching pairs we derive p-value for the comparison of distribution with respect to the distances.

The optimal, non-bipartite matching algorithm is written in C by Ed Rothberg [4], and implements H. Gabow's N-cubed weighting matching algorithm [2]. The algorithm strives to maximize a total benefit, defined as β_{ij}, as opposed to striving to minimize a total distance δ_{ij} [3]. In effect, the total benefit is defined as

$$\beta_{ij} = \max_{b,c}(\delta_{bc}) - \delta_{ij}. \tag{3.4}$$

3.4 Significant Figures

The optimal, non-bipartite algorithm implemented in C accepts only integers as input values for distances between two points. If the Mahalanobis distance vector entered into the code contains decimal places, the C function will remove the additional figures in the conversion from double precision floating-point numbers to integers. As a result, the optimal number of non-matching pairs may be inaccurate due to the potential of failing to account for significant figures contained within the distance sums. One solution to avoiding this error is multiplying the Mahalanobis distance vector by 10^N, for some sufficiently large N, thereby ensuring that the conversion to integer format encompasses these significant figures that were not previously considered in the former number conversion process.

At some point, when all distances are different from each other, or reflect the most accurate number of significant figures obtainable, the solution is most rigorous and may not be improved for any greater N value. The upper bound for this improvement is the ability of R to express double precision floating-point numbers, so in the worst case where two distances are dissimilar but only noticeable after more than 16 significant figures, this method may not provide sufficiently accurate results. The R function created for this project requires an input value for N, so that the user may particularly specify the degree of accuracy she wishes to obtain from the resulting number of optimal non-matching pairs.

3.5 Exact Null Distribution

Considering the location data of the ribosomal proteins Y, the null distribution for A_1, the total number of crossed-matched pairs, is represented by

$$Pr(A_1 = a_1|Y) = \frac{2^{a_1}\, I!}{\binom{N}{n}\, a_0!\, a_1!\, a_2!} = \pi_{a_1}, \tag{3.5}$$

where A_k is defined as the number of pairs with k treated subjects, $A_0 + A_1 + A_2 = I$ is the number of total pairs for the entire set of N data points, and n is the number of cofactor phenotype ribosomal proteins [3].

As proven in [3], the null distribution for A_1 converges to the normal distribution, where

$$z = \frac{A_1 - E(A_1)}{\sqrt{Var(A_1)}}. \tag{3.6}$$

3.6 Accuracy of p-value

While it has been shown that the distribution of the cross-match statistic converges to the normal distribution, we also chose to calculate a more exact p-value than the approximation provided by using the normality assumption. In order to calculate this p-value, we performed a permutation test where we randomly assigned each data point to a particular group. We then ran our cross-match tests in order to see what the distribution of the cross-matches would be like if the distribution was random. We repeated this process 5,000 times in order to see how many cross-matches would occur.

4 Results

After computing the Mahalanobis distance between all possible pairs of ribosomal proteins, the optimal, non-bipartite sorting algorithm determined that there are 17 non-matching pairs in the closest within-pair distance pairing. Figure 4.1 offers a visual plot showing the optimal connections between all pairs, where the cyan lines represent the connections between phenotype and non-phenotype proteins. Since there are 76 proteins in the sample, the the maximum possible amount for non-matching pairs is $76/2 = 38$. In effect, roughly 45% of optimal pairings are non-matching.

Alternately, to emphasize the non-matched pairs between the two categories, Figure 4.2 displays the optimal within-distance pairing while only showing the connections between phenotypes to non-phenotypes.

In this computation, the number of significant digits for distance were specified as $N = 3$, which is equivalent to multiplying the Mahalanobis distance vector by 10^3. The same conclusion was also obtained for $N = 5$, and therefore $N = 3$ is sufficiently large enough to reflect the most accurate optimal pairing of phenotypes to non-phenotypes in this case.

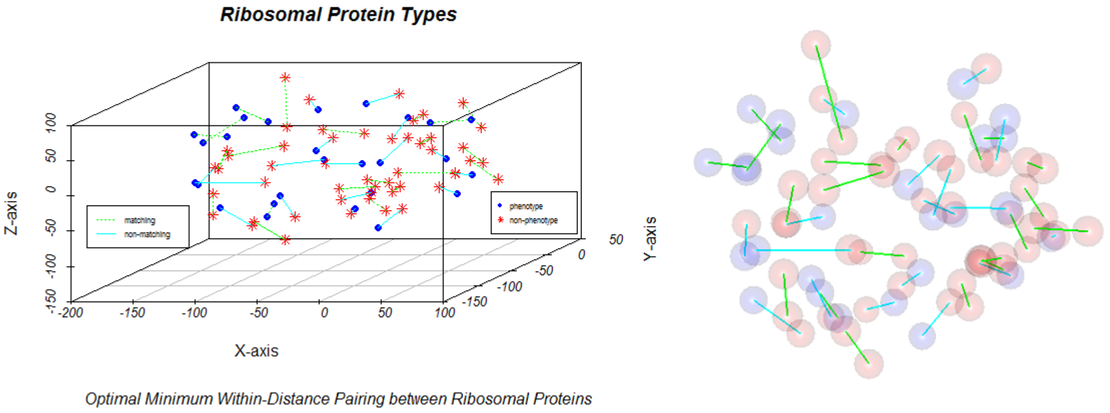

Figure 4.1: Optimal within-distance pairing between ribosomal protein types, all matching.

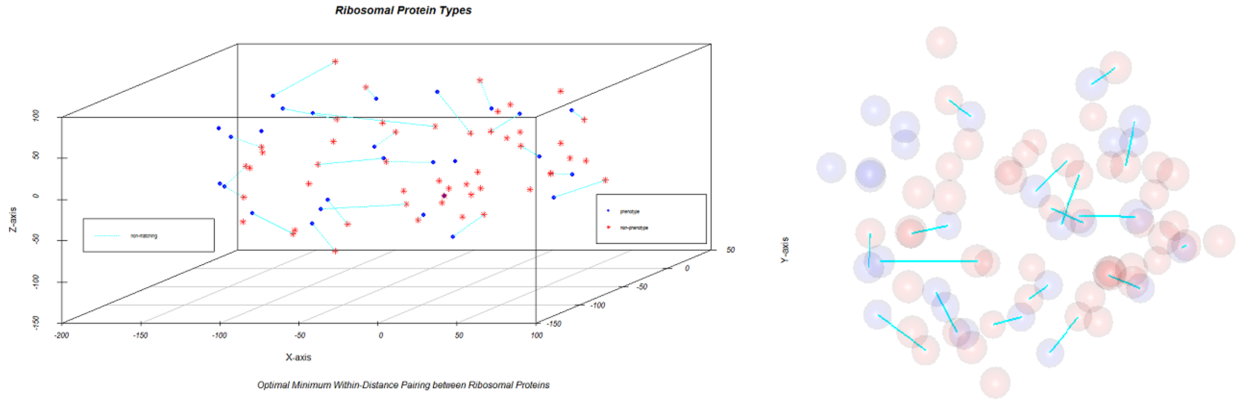

Figure 4.2: Optimal within-distance pairing between ribosomal protein types, non-matching.

Let the null hypothesis claim that the distribution of three-dimensional positioning is the same for phenotype and non-phenotype proteins. Computing a p-value using the fact that the conditional distribution of A_1 converges to the normal distribution [3] (see Numerical Methods section), it follows that $p = 0.821447$. Consequently, the null hypothesis is failed to be rejected, and therefore the distributions of three-dimensional positions are statistically similar for both groups.

Determining whether substantial variance is present within the two groups, the Euclidean distance is computed in place of the Mahalanobis distance and the deviation from the above result is noted. Using the 3×3 identity matrix rather than the variance-covariance matrix, the 2-norm is determined for all possible within-pair distances and optimal, non-bipartite matching is utilized to obtain the minimum within-distance sum. The number of non-matching pairs is equal to 21, which is slightly greater than the above Mahalanobis result. Computing a p-value from these pairings, it follows that $p = 0.236081$. Again, the null hypothesis has failed to be rejected at an alpha value of 0.05, and therefore the distributions of

Table 4.1: Summary statistics for non-matching pairs.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.00	15.00	17.00	17.61	19.00	25.00

three-dimensional positions are still statistically similar even not taking into account inter-group variance. In this case, the number of significant digits is expressed to four significant figures in the integer conversion.

Finally, performing a permutation test in which randomly assigned data in the set are identified as either phenotype or non-phenotype, consistent with the proportion present in the original set, the number of optimal non-matching pairs was computed at each iteration. This process was repeated $5,000$ times to note the distribution of non-matching pairs from each instance. Table 4.1 shows the summary statistics of these non-matching pairs in this experiment, and one should confidently notice that the above number, 17, is well within the middle range.

Figure 4.3 is the histogram for our permutation test, which allowed us to calculate a more exact p-value, with a violin and box plot to reinforce the above numerical conclusion.

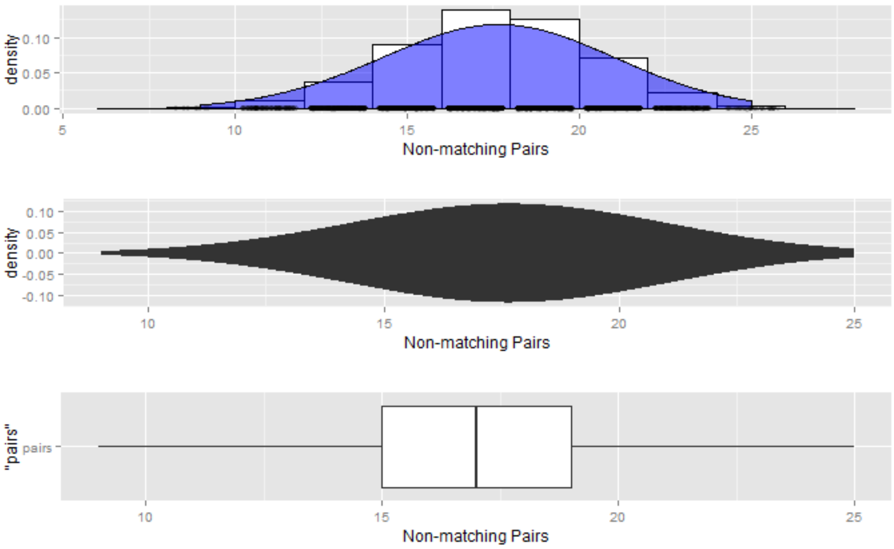

Figure 4.3: Histogram for non-matching pairs overlaid with kernel density curve, with violin and box plots.

5 Conclusions

After computing the Mahalanobis distance between each protein pair within the data set, it was determined by the optimal, non-bipartite sorting algortihm that there are 17 non-matching pairs of ribosomal proteins matching phenotype to non-phenotype traits. Since the data set contains 76 elements, the maximum possible non-matching pairs is $76/2 = 38$ pairs, and therefore one should notice the high proportion of non-matching pairs $A_1 = 17$.

Therefore, a null hypothesis may be constructed claiming that the distributions of phenotype to non-phenotype positions are similarly adjacent. Computing a p-value based on the above non-matching pairs, equal to 0.821447, one may reject the null hypothesis at a standard alpha level equal to 0.05. That is, one may fail to reject the null hypothesis, and conclude that the distributions comparing adjacency between phenotype and non-phenotype groups are statistically equivalent. Further, this conclusion is supported by the high proportion of non-matching pairs determined above.

Acknowledgments

These results were obtained as part of the REU Site: Interdisciplinary Program in High Performance Computing (`www.umbc.edu/hpcreu`) in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County (UMBC) in Summer 2013. This program is funded jointly by the National Science Foundation and the National Security Agency (NSF grant no. DMS–1156976), with additional support from UMBC, the Department of Mathematics and Statistics, the Center for Interdisciplinary Research and Consulting (CIRC), and the UMBC High Performance Computing Facility (HPCF). HPCF (`www.umbc.edu/hpcf`) is supported by the National Science Foundation through the MRI program (grant nos. CNS–0821258 and CNS–1228778) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from UMBC. Co-author Jordan Ramsey was supported, in part, by the UMBC National Security Agency (NSA) Scholars Program though a contract with the NSA. Graduate RAs Xuan Huang and Andrew Raim were supported by UMBC as HPCF RAs.

References

[1] Erik B. Erhardt, Edward J. Bedrick, and Ronald M. Schrader. Lecture notes for Advanced Data Analysis 2 (ADA2). University of New Mexico, Spring 2014, `http://statacumen.com/teach/ADA2/ADA2_notes_S14.pdf`, accessed September 03, 2013.

[2] H. Gabow. Implementation of algorithms for maximum matching on nonbipartite graphs. Ph.D. thesis, Stanford University, 1973.

[3] Paul R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Statist. Soc. B*, 67:515–530, 2005.

[4] Ed Rothberg. MATHPROG: Solver for the maximum weight matching problem, 1999. `http://elib.zib.de/pub/Packages/mathprog/matching/weighted/index.html`, accessed September 03, 2013.