Hemoglobin Response to Higher Order Gene Interactions

A Spectral Analysis Approach

MSRI-UP 2018 Technical Report

Lillian González Albino[1], Rosa Garza[2], and Sylvia Akueze Nwakanma[3]

[1]University of Puerto Rico, Río Piedras
[2]California State University, Monterey Bay
[3]Pomona College

July 2018

Abstract

Mutations pave the way for evolution, which is why with the growing amount of genomics data we need efficient and effective ways to analyze interactions of gene mutations. In this paper, spectral analysis is used to orthogonally decompose a genomic data set and analyze higher order interactions between mutations. This approach offers insight into the effects that certain gene mutations groups have on hemoglobin while accounting for redundant information. A comparison was performed between our methodology and classic multilinear regression models to highlight some of the advantages that our approach has in identifying significant mutation interactions. The results from our research can be extended to analyses of the variations observed in other phenotypes.

Keywords: Spectral Analysis, Single Nucleotide Polymorphism, Gene Interactions, Higher Order Interactions, Orthogonal Decomposition, Data Vector, High Dimensional Data Analysis

Contents

Acknowledgements

We would first like to thank our primary mentors on this project, Dr. David Uminsky and Dr. Mario Bañuelos for their persistent enthusiasm, patience, and late-night doughnuts. Their energy and deep wealth of knowledge always left us fully engaged in our work. We also extend our gratitude to the other Teaching Assistants in the program, Dr. Paul Ignacio and Joanna Navarro, for their continued assistance throughout this journey. We thank Dr. Mercedes Franco for being a mentor, an enthusiastic practice audience, and a friend. Her unyielding belief in our intellectual capabilities has been a guiding light. Last but not least, we would like to thank our sponsors at the Alfred P. Sloan Foundation (Sloan Grant: G-2017-9876), the National Security Agency (NSA Grant: H98230-18-1-0008), the National Science Foundation (NSF Grant: DMS-1659138), and the Mathematical Sciences Research Institute at Berkeley for supporting the endeavours of programs like MSRI-UP.

1 Introduction

Hemoglobin are important proteins in red blood cells which are responsible for transporting oxygen throughout the body [17]. When people travel or visit places at high altitudes (approximately 4,500 meters above sea level) their hemoglobin levels tend to go up to compensate for the decrease in oxygen [5, 4]. For people that live at high altitudes, however, this is not always the case. Their bodies adapt to the decrease in oxygen in different ways that do not involve raising hemoglobin levels. These adaptations are made possible by certain gene mutations [3, 18]. Given the vast amount of data on different gene mutations, predicting the leading causes of certain phenotypes is a non-trivial task. Our project specifically examines the hemoglobin levels of 146 individuals living in high altitudes and the presence of thirteen hand-picked gene mutations possibly associated with hemoglobin. These mutations were chosen by Dr. Emilia Huertas-Sánchez for her research in adaptation to high altitudes. We aim to identify which interactions of these mutations correlate significantly with high or low hemoglobin levels.

Classical methods for analyzing data sets like ours involve assuming an underlying structure to the data first and then fitting data points to a regression model that supports the assumptions made. One evident limitation of this method is having prior knowledge about the data set, or having to make structural assumptions based on a data summary. Our analysis is exploratory; we let the data reveal its own underlying structures by using spectral analysis to decompose the data into principle components.

2 Preliminaries

2.1 Vocabulary

Below are some key terms used throughout the paper to describe our application of spectral analysis.

k-Grouping: Refers to the subset of the data in which every observation within the subset has exactly k mutations (non-mutations).

Coalition: Is a set of mutations (or non-mutations) in a particular k-grouping.

i^{th} **Order**: The i^{th} spectral decomposition of a k-grouping.

i^{th} **Order effect**: The effect that those mutations (or non-mutations) have on hemoglobin levels in observations with exactly k mutations.

Pure effect: The pure effect of a coalition on hemoglobin levels is synonymous with the i^{th} order effect of that coalition.

5 Results

5.1 Exploratory Data Analysis

As an initial data exploration, we created histograms of the number of gene mutations people had in the sample for both the OR and the AND reduction of the data. With the histograms, we were able to study the overall variance in the number of gene mutations that individuals had (see Figure 5 below).

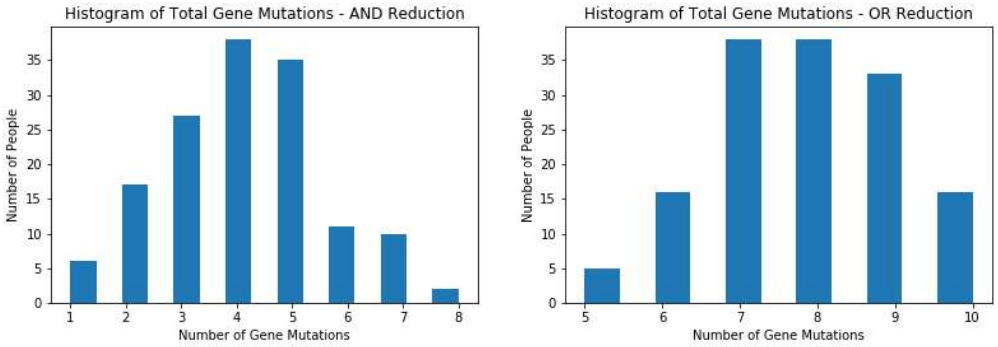

Figure 5: Histograms of the total number of gene mutations with the number of people who specifically had a particular amount of gene mutations in the AND reduction of the data (left) and the OR reduction (right)

In the OR reduction, no individual had fewer than 5 mutations total or more than 10 mutations total. Most people had 7, 8, and 9 mutations out of the 13. As one would expect, there is more variance in the number of mutations people have under the AND reduction since there is a stricter constraint on assigning 1's. One of the questions we explore is how significant having a mutation in both mutation copies is compared to having only one copy.

We also created a scatter plot for both data set reductions where the y axis is the hemoglobin levels of each individuals and the x axis is the individuals in no particular order (refer to Figure 6 below). There does not appear to be any distinguishable pattern in this plot. We would not even be able to fit the data to any regression model since we cannot assume any underlying structure. Spectral analysis is ideal for a more in depth exploratory analysis.

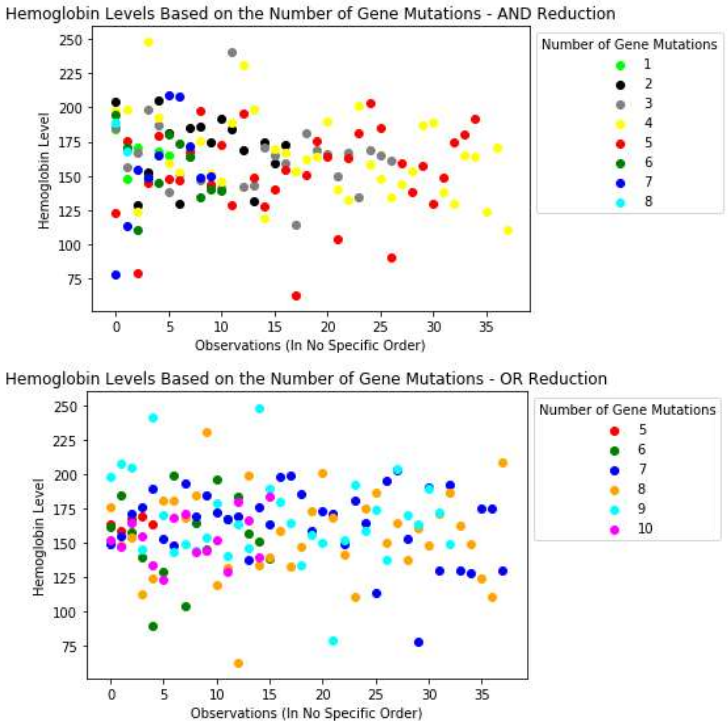

Figure 6: Scatter plot of the hemoglobin levels for each individual in the AND (top) and OR (bottom) data set. The colors correspond to the total number of mutations each individual has in the respective data set reduction

5.2 Multilinear Regression

We performed a multilinear regression on the largest subset of our cleaned data, the $f^{(13,4)}$ data vector in the AND reduction with 37 observations. We used two linear models: in the first, we regressed hemoglobin against all 13 mutations. In the second, we regressed hemoglobin against every combination of the 13 mutations that occurred in the data set. In both, the adjusted R^2 values are below 10 percent as reported in Figure 7 and 8 below. In fact, the first model performed so poorly that the adjusted R^2 is below 0. Although the second model is an improvement, it shows that only about 10 percent of the variation observed in hemoglobin can be explained by a linear combination of the mutations. We note that the coefficients for a is persistently high in both multilinear models, and fluctuates in sign when paired with other mutations. This is consistent with the frequency of mutation a in the spectral analysis. While there is some overlap in the information extracted from the multilinear regression and from the spectral analysis, the latter approach produces more gainful insight as demonstrated in subsequent sub-sections.

```
`lm(formula = Hemo ~ a + c + d + e + f + g + h + j + m)`

Residuals:
    Min      1Q  Median      3Q     Max
-71.208 -19.487  -3.713  21.886  75.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    73.39      59.18   1.240   0.2248
`   a           31.53      22.75   1.386   0.1763 `
    c           11.66      20.95   0.557   0.5820
    d           17.62      19.35   0.911   0.3700
    e           21.27      17.27   1.232   0.2280
`   f           31.13      23.89   1.303   0.2027  `
`   g           32.72      18.41   1.778   0.0860 .`
`   h           32.76      18.40   1.780   0.0856 .`
    j           17.02      18.51   0.919   0.3655
    m         -16.5631    25.6966  -0.645    0.524

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 35.23 on 29 degrees of freedom
Multiple R-squared:  0.1595,    Adjusted R-squared:  `-0.07235`
F-statistic: 0.688 on 8 and 29 DF,  p-value: 0.6987
```

Figure 7: Multilinear regression of hemoglobin on the 13 mutations in $f^{(13,4)}$ in the AND reduction of the data set.

```
lm(formula = Hemo ~ a * c * d * e * f * g * h * j * l * m)

Residuals:
   Min     1Q Median     3Q    Max
 -56.5    0.0    0.0    0.0   56.5

Coefficients: (997 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         450.50     326.05   1.382   0.1945
`   a               -41.50     112.27  -0.370   0.7187 `
    c              **-245.00**  232.92  -1.052   0.3154
    d              **-252.75**  239.30  -1.056   0.3135
    e              **-292.75**  197.94  -1.479   0.1672
    f               **5.00**    182.83   0.027   0.9787
    g               **-4.50**   130.34  -0.035   0.9731
    h                -57.75     118.21  -0.489   0.6348
    j                 45.00      46.82   0.961   0.3571
    l                -82.00      46.82  -1.751   0.1077
`   a:d               65.25      95.09   0.686   0.5068`
    c:d              112.00     134.47   0.833   0.4226
`   a:e               46.50      90.66   0.513   0.6182 `
    c:e              214.25     122.76   1.745   0.1088
    d:e              106.75      72.15   1.480   0.1671
`   a:f              -27.75     105.99  -0.262   0.7983 `
    c:f              -30.75     100.69  -0.305   0.7658
    d:f              111.75      97.93   1.141   0.2780
`   a:g               58.00      52.34   1.108   0.2915 `
    c:g              -50.50      93.64  -0.539   0.6004
    d:g              106.00      57.34   1.849   0.0915 .
    e:g               26.00      84.40   0.308   0.7638
    f:g              -39.75      68.25  -0.582   0.5720
    c:h               44.75     108.54   0.412   0.6881
    d:h               67.25      54.90   1.225   0.2462
`   a:j               13.00      57.34   0.227   0.8248 `
    c:l              142.50      61.93   2.301   0.0420 *


Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.11 on 11 degrees of freedom
Multiple R-squared:  0.7185,    Adjusted R-squared:  `0.05313`
F-statistic:  1.08 on 26 and 11 DF,  p-value: 0.4686
```

Figure 8: Multilinear regression of hemoglobin on every combination of the 13 mutations that occur in $f^{(13,4)}$ in the AND reduction of the data set.

5.3 Methodology Comparison: Keeping Zeros vs. Imputing the Zeros in the Data Vector

We consider noise to be high or low values on our f_i vectors that correspond to coalitions of mutations that do not occur in the raw data; there are two types of noise: the first is when one or more mutations in that coalition does not occur in the data at all (we will call this type I noise) and the second is when all mutations as individuals occur in the data but the combination itself does not (we will call this type II noise). After plotting both the data vector with zeros

on entries where there where no individuals with that coalition of mutations and the imputed data vector, we analyzed the highest and lowest peaks on each graph for each order. What we found is that the noise cause by leaving the zeros in the data vector is very high and carries over for high orders as well. However, for the imputed data vector, although there was noise, it was significantly less and it did not persist in high orders. In the next table we will show the high and low peaks on each order for individuals that have exactly 6 mutations and using the OR reduction data set . It is important to note that although we only show this table for this grouping, it is true for the rest of the groupings as well on both reduction of the data sets.

Order	Highest Peaks		Lowest Peaks	
	Non-Imputed	Imputed	Non-Imputed	Imputed
First	C, A, D	E, M, G	**B**,**K**,**I**	H,D
Second	AC, **BK**, **IK**, **BI**, CD, AD, DF, **KL**	EH, DM, FM, CG, *JL*	**CI**, **BC**, **CK**, **AI**, **AK**, **AB**, *GL*, **DI**, **DK**, **BD**, CJ	DH, AH, *EM*, FH, DF, GJ, CH
Third	ACD, CDF, ADF, ACF, **BCK**, **CIK**, **BCI**, **ABI**, **AIK**	AEH, DFM, CEH, AFM, ADM, CFG, CFM, ACG, CDM	**BIK**, **IKL**, **BKL**, **BIL**, **ACI**, **ABL**	DFH, ADH, **BEH**, **EHK**, **EHI**, CDH
Fourth	ACDF, **BIKL**, **BIJK**, CDGM, ACEF	ADFM, CDFM, ACEH, ACFG, ACFM, ACDM	*AFHM*, *CEGM*, **BCIK**, **ABIK**, *ACEM*, *ACGL*, *CDGJ*, *ACDJ*	ADFH, CDFH, ACDH, CGHM, AGHM, **AEHK**, **AEHI**
Fifth	CDFGM, ADFHJ, ACEGH, ACDGM	ACDFM, ADFGM, ACEGH, CDFGM, ACFGM, ACEFH	**ACFHM**, **BIJKL**, **ACDGJ**	ACDFH, ACGHM, CDEFG
Sixth	ACGHJM, ACDFGM, ADFGHJ	ACDFGM, ACEFGH, **ACDFHK**	*ACDFHM*, *ACFGHM*, *CDEFGH*	ACDFHL, ACDEFG, ACDGHM

Table 1: Table of high and low peaks for non-imputed and imputed data vectors for individuals with exactly 6 mutations. In bold we marked type I noise. In italics we marked type II noise.

Like we explained in Section 3.4, high and low values on our Mallow transformed data vectors mean that the corresponding coalition of mutations in that k-grouping of mutations contribute to higher and lower than average levels of hemoglobin respectively on individuals with exactly k mutations and that have that specific coalition of mutations. When we graph these vectors, the highest and lowest peaks represent these coalitions with higher and lower than average effects on hemoglobin levels, which is why we look into the highest and lowest peaks or values of each data vector.

If we recall Example **??**, when we did not impute the data vectors, the average hemoglobin for $f^{(4,2)}$ was 80.8 and in comparison, all non-zero valued

coalitions of mutations in $f^{(4,2)}$ seemed to have a high effect on hemoglobin. When we imputed our data vectors, the average of all non-zero valued coalitions was 168.48 and we could see that the coalitions that were above average (BD) were the ones that contributed to higher hemoglobin and the coalitions that were under the average (AD and BC) contributed to lower hemoglobin. We corroborate this in Example 3.1 where from $f_2^{(4,2)}$ we see that coalitions AC and BD contribute the most to high hemoglobin and AD and BC contribute the most to low hemoglobin. If we were to summarize this example in a similar table as Table 1 we would get the following:

Order	Highest Peaks		Lowest Peaks	
	Non-Imputed	Imputed	Non-Imputed	Imputed
First	*B*, D	A, C	A, C	*B*, D
Second	AD, BC	*AB*, *CD*	*AB*, *CD*	AD, BC

Table 2: Table for high and low peaks for non-imputed and imputed data vectors for Example 3.2. In italics we marked type II noise.

Using this table we can see that the imputed method detected that AD and BC contribute to low levels of hemoglobin. It also shows AC as a high contributor to hemoglobin levels but we marked it in italics because that coalition does not occur in the data even though A and C do occur. From here on out we will continue to use only imputed data vectors for the rest of our analysis.

5.4 AND Reduction

5.4.1 Comparison of Mallow's Method on Raw Data Vectors and Imputed Data Vectors

To validate why we do not apply Mallow's method on the raw data and instead apply it to the decomposed vectors, we compare the results of Mallow's method on the raw data vectors and on the imputed data vectors. We focus on the first order space to see the difference in the relative significance of individual mutations.