The Depths of the Human Genome:

Data Mining and Enhancers

Laura Colbran Jimin Yoo

Dave Musicant
CS324 Data Mining

3.16.15

1 Introduction

Biology is one scientific field that has been revolutionized by the advent of fast computers and big data. The realm of genetics and genomics has particularly grown with the expansion of fast DNA sequencing techniques and lends itself to computational methods of analysis. DNA is the blueprint for building an organism, and as such, there is a lot of information contained in its sequence. Every cell in our bodies contains a full copy of the information to make any protein the body might need. This information is stored as a double-stranded sequence composed from 4 base molecules. The genome is packaged into separate chromosomes within the cell, and each chromosome contains hundreds of genes, specific sequences that each code for a particular protein (Figure 1). The information is 'read' by proteins called RNA-polymerases, which bind to the gene and transcribe its genetic sequence, which is then translated into protein by other molecules based on the order of those 4 base molecules. Different cell types have unique sets of proteins because they express a unique subset of the genes on the genome.

The human genome has about 3 billion base pairs, a quantity that is impossible for humans to analyze without the help of computers. However, only 2% of the human genome is contained within genes. For a long time scientists thought the other 98% was just redundancies left over from millennia of evolution, but there has been a lot of recent evidence that sequences in non-protein-coding regions are involved in regulating which genes are expressed. The fact that much of the non-coding region seems to have a function after all means that alterations to the sequence of base pairs have the potential to alter the overall function of the organism even when the alteration isn't in a gene, and scientists have accordingly begun to focus on these noncoding regions in their work.

Enhancers are a particular subset of those regulatory sequences that interact with a target gene (or genes) to increase expression of those genes. Researchers have found evidence that mutation in the sequences in

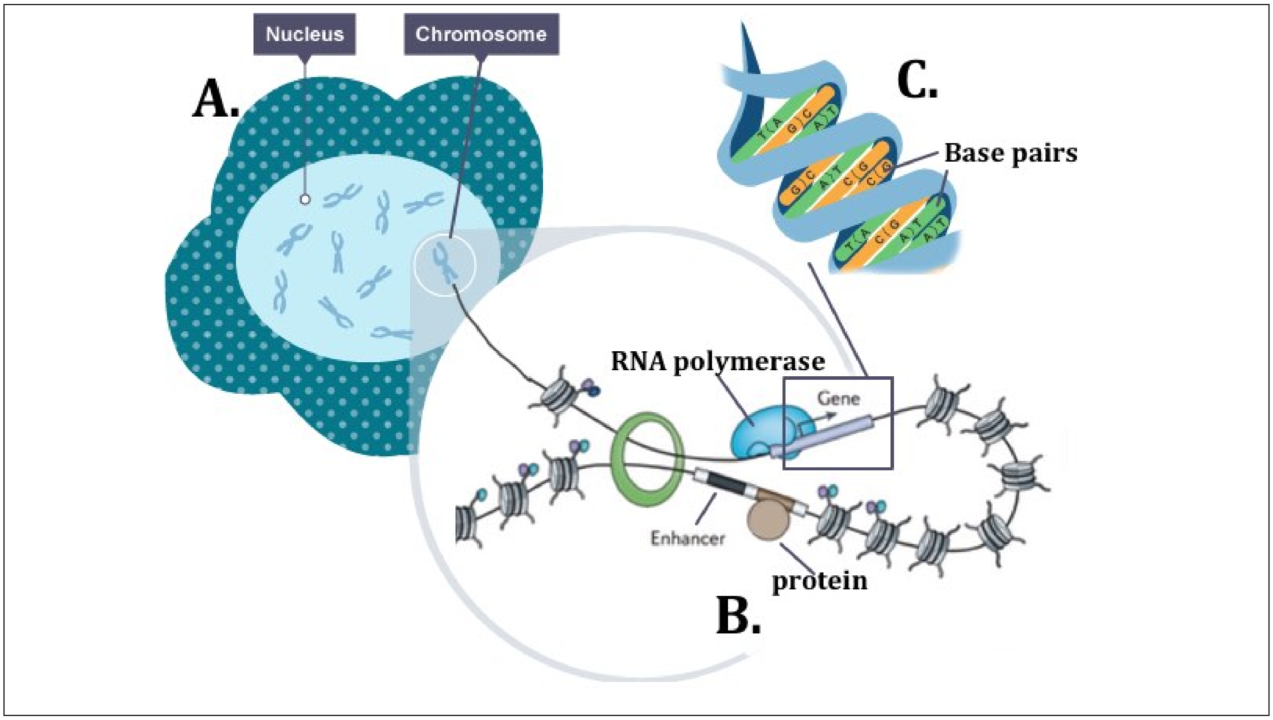

Figure 1: A) Within the nucleus of a cell, DNA is packaged into chromosomes. Each chromosome is one single strand of DNA. B) Genes are sections of DNA sequence that encode information for building proteins. Enhancers are different sections that interact with genes to increase expression of those genes. Some have binding sites for regulatory proteins. Multiple enhancers can affect the same gene, and the a single enhancer can have multiple targets. C) RNA polymerase reads the sequence of pairs of base molecules of the gene. Adapted from Shlyueva 2014[3] and BBC Bitesize.[2]

known enhancers are involved in genetic diseases like cystic fibrosis and in many cancers. Because it is relatively easy these days to find out the DNA sequence of individuals with a particular disease, the big problem in researching has been finding out the function of particular regions of the genome where potentially relevant mutations have been observed. Enhancers are difficult to find and identify, mostly because there is no characteristic that's universal for all the ones we've experimentally verified. Enhancers are no set distance away from their target gene; they can be thousands of base pairs away, in the middle of another gene, or even on a completely different chromosome. Additionally, they don't necessarily have only one target gene. Many of them have sequences that proteins can bind to, but that is not very helpful for identifying them because not all bind to the same protein.

There has been a big push recently to try and find new ways of identifying potential enhancers. The FANTOM5 consortium is the most recent in a long line of attempts, and has published by far the largest set of putative enhancers.[1] Their goal in this project was to identify human enhancers that are active in hundreds of different cell types, as well as in cells of individuals with particular genetics diseases. The published data includes over 38,000 putative enhancers, as well as expression information for 800 different varieties of cell. Our final project was to apply some of the data mining techniques we learned in class this term to this dataset, to see if we could glean any new information from it. Particularly, we hoped to cluster the enhancers by their location and number of tissues of activity, and use association rules to potentially find patterns in which tissues they were active in.

2 Results

We used several different types of data mining in R to discover patterns in the FANTOM5 data. The correlation between length of the enhancer and the number of cell types it was active in was r = 0.33, which was not a convincing enough relationship for us to commit to more intensive study of the enhancer lengths. We did, however, look to see if there were any patterns in the locations of the enhancers within the genome.

First ,we tried the k-means clustering based on each enhancer's chromosome location ("chr"), the starting and ending locations of the enhancer within the chromosome, the enhancer's length ("length"), and the enhancer's level of activity measured by the number of tissue types an enhancer is active in ("activity"). In our first investigation of the clusters, we decided that the starting and ending locations were not interesting factors, because they were evenly distributed among all clusters. Thus, we took out the starting and ending locations from the analysis. After looking at the plot of total clustering error vs. k, we decided that the "knee" of the curve is at $k = 7$ (Figure 2).Thus, we clustered enhancers into 7 groups. However, we found that the clusters did not tell much interesting information. The seven clusters all had chromosome location centers around chromosome 9. The clusters showed most distinctively divided in terms of enhancer length, but showed non-distinctive patterns in the activity level (Figure 3). Since activity level difference was the factor we were much more interested in than length difference, we concluded there is no drastic difference among clusters to consider them significant.

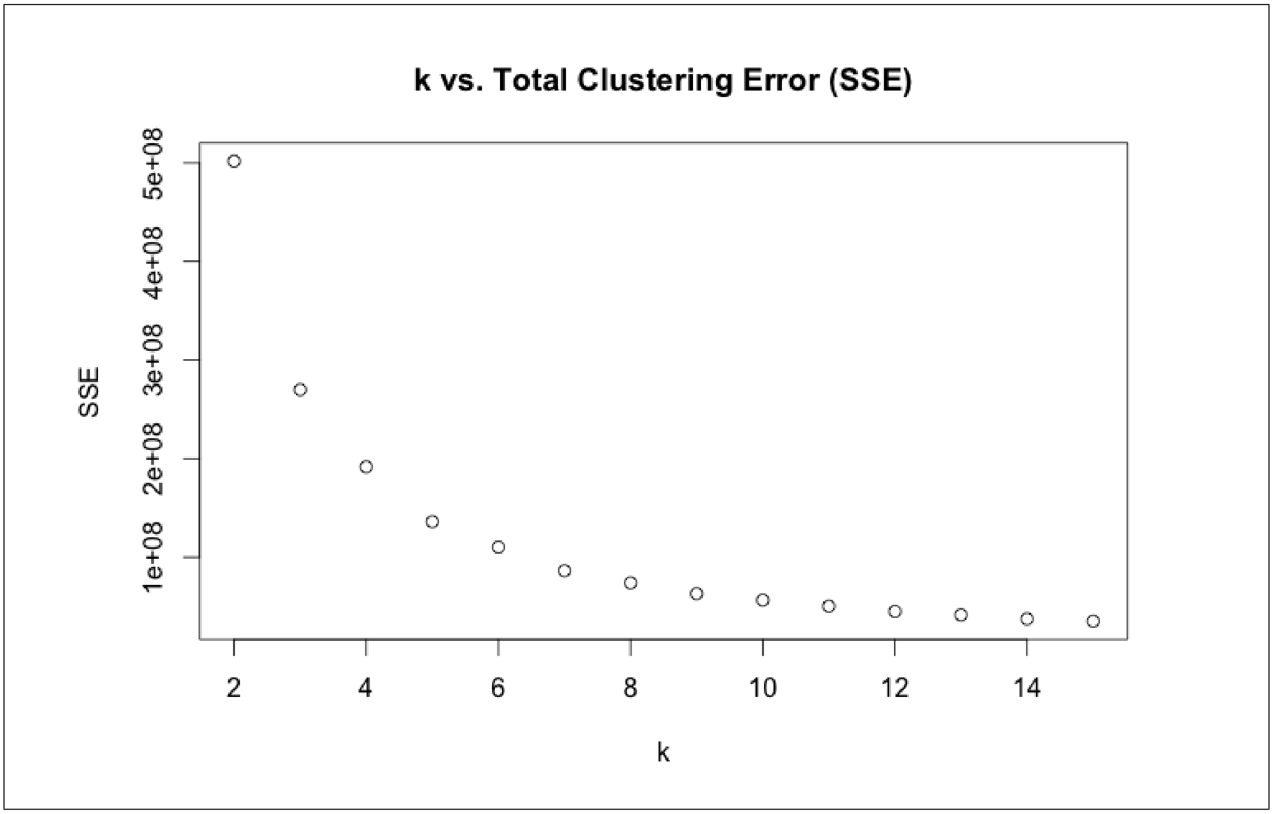

Figure 2: Total clustering error for each choice of k.

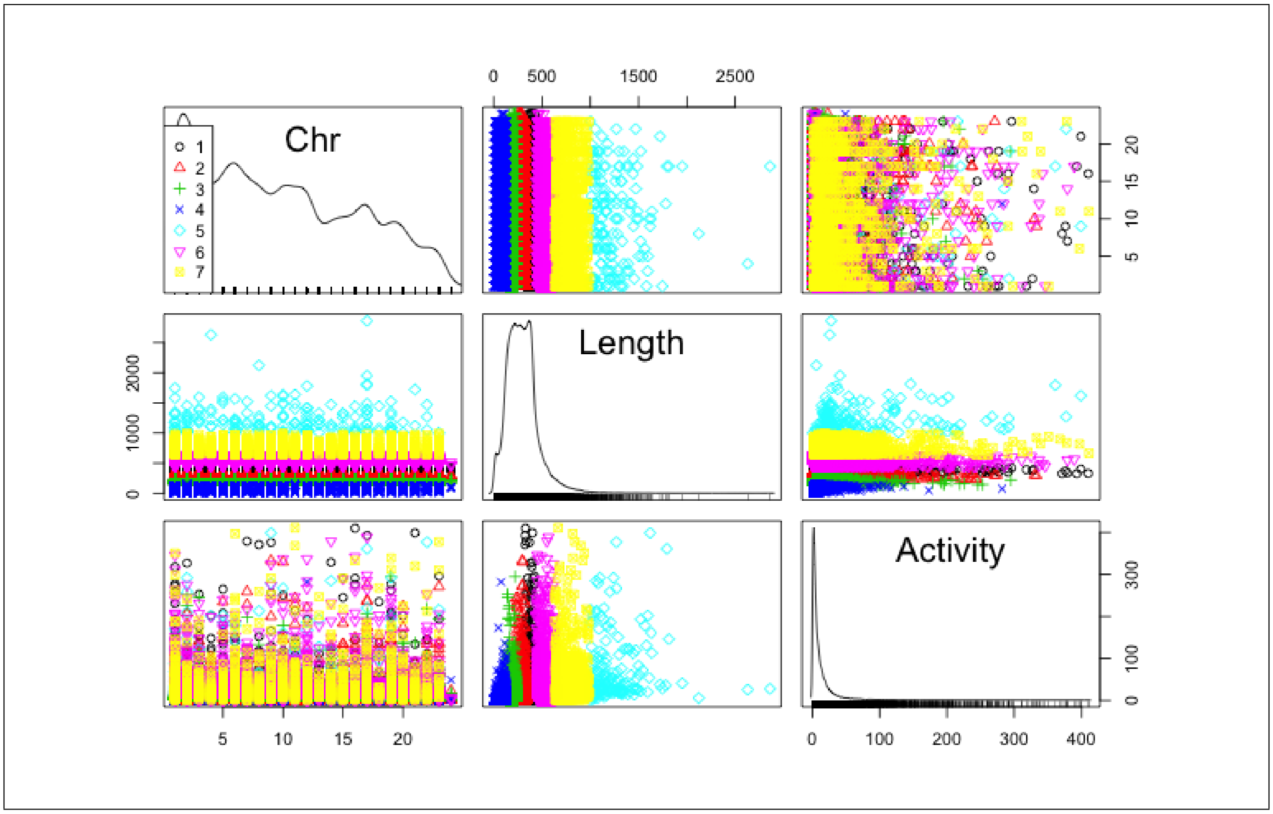

Figure 3: Distribution of chromosome location, length, and activity level (measured by number of tissue types an enhancer is active in) for the seven clusters.

Considering that uninteresting results from K-means clustering may be due to non-globular clusters, we also tried hierarchical clustering technique. Andersson et al. conducted hierarchical clustering on the data set when they published their paper, but the only result they reported was a cluster of enhancers that were ubiquitous, i.e. active in many tissue types [1]. We ran into problems, however, when we tried to cluster the entire data set. None of the computers we were using had the power to construct the clusters, and ended up either crashing or tying themselves in knots. Therefore, we took a random sample of 5000 enhancers from the full set, and examined the order height of merge of those clusters to see if they mirrored the trend observed by Andersson. There was a cluster of points with high activity that merged with the rest of the clusters at a dramatically larger height than any of the others, which suggested that our sample also had a separate cluster of highly active enhancers. There didn't seem to be any other particular clusters that stood out from the rest, but without the computing power to examine them more closely, we can't draw any more conclusions.

Besides location and activity of the enhancers, we were curious to see if there were any predictions for the tissues they would be active in. To test this, we first attempted to run *a priori* association rules for the entire dataset. Our goal was to find association rules that would indicate if an enhancer is active in certain tissue(s), it is also likely to be active in the "associated" tissue. After some experimentation, we concluded it was impractical to do that with so many enhancers and so many possibilities for frequent itemsets. To make the amount of memory involved more manageable, we extracted a chunk of the data for enhancers that were active in various Lung cells. This gave us 4120 enhancers that were active in at least one of the following sample types: Adult Lung (AL), Fetal Lung (FL), Right Lower Lobe (RLL), Lung Adenocarcinoma (LAC), Large Cell Carcinoma (LCC), Small Cell Carcinoma, (SmCC) and Squamous Cell Carcinoma (SqCC). The association rules with the five highest interest ("lift") can be found in Table 1.

Association Rule	Support	Confidence	Lift
{SmCC, SqCC, LAC, AL, FL} ⇒ {LCC}	0.009	0.814	3.489
{SmCC, SqCC, AL, FL} ⇒ {LCC}	0.012	0.800	3.429
{LCC, LAC,FL} ⇒ {AL}	0.014	0.906	2.464
{SqCC, LCC, LAC, FL} ⇒ {AL}	0.001	0.889	2.417
{SmCC, LCC, LAC, FL} ⇒ {AL}	0.011	0.887	2.411

Table 1: Five association rules by highest interest (lift)

The association with the highest confidence (0.906) was {LCC, LAC, FL} ⇒ {AL}. This also had the reasonably high lift of 2.46. Many of the rules generated had FL associated with AL, plus some extra cell types on either side. This makes sense, since both are associated with healthy lungs, just at different developmental stages. However, the more interesting enhancers are probably the ones that don't follow that pattern. Those enhancers could play a role in activating genes responsible for development specifically, not for overall lung function.

None of the rules generated had high support. This means that while the rules were consistent, there just weren't that many enhancers that the rules were true for. Put another way, there weren't any association patterns that occurred in anything other than very small numbers. This serves to demonstrate the high specificity and variety in the function of enhancers. The highest support seen was 0.028, which is around 100 enhancers with the pattern {LCC,FL}⇒{AL}. Out of the original 38,000 enhancers that's negligible. If this pattern holds throughout the whole dataset, that suggests that there are few truly multipurpose enhancers. Rather, there exists a specific group of enhancers for each function.

3 Conclusions

We utilized kmeans clustering and hierarchical clustering techniques to find "good" clusters of enhancers. Kmeans clustering revealed clusters were not significantly different in terms of chromosome location, starting and ending locations within the chromosome, and activity level, which we suspect as due to non-globular clusters. The biggest challenge of hierarchical clustering came from interpreting the result, as it was almost impossible to create a visible dendrogram even with a sample of 5000 enhancers. We also applied association rules to a sample of 4120 enhancers, but concluded that results were not of significant interest as all the association rules had very small support.

Our results suggest that big, wide-reaching data mining methods like the ones we learned this term are less than ideal for looking at enhancers as a class of regulatory sequence. While patterns do exist within the data, they are limited to very small numbers of enhancers, and are often not drastically different enough for algorithms like kmeans clustering to yield useful results. Our computational limitations were the main culprit for us being unable to examine hierarchical clustering in enough detail to learn anything new, but we were able to see similar results to those reported in the original FANTOM5 paper by Andersson et al.

While we didn't find anything particularly exciting, our results are a good illustration of the complexity behind genetic regulation. Better targeted mining techniques and more computing power might still find new patterns in the raw enhancer data. Besides, none of this affects the potential use that can be gotten out of this data set in screening for functional mutations in noncoding regions, or in studying specific diseases.

References

[1] Andersson, R. et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature. 507, 455-468.

[2] BBC Bitesize. (2015). DNA and cell division. http://www.bbc.co.uk/education/guides/zvb7hyc/revision/1

[3] Shlyueva, D. et al. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews. 15, 272-286.