Application of Data Mining Techniques to Human Genome Data (HapMap project)

Alex Rakitin (Natera)

April 2, 2014

GA SF Data Science Course

Introduction

- I am working at a company which deals with human DNA (Natera)
- For the project I wanted to take something related to genetics, *e.g.* prediction of certain human traits based on genetic information
- For this purpose I needed a dataset collected from different people together with their genetic data and at least one phenotype* trait

**Phenotype* is a collection of all observable features and traits which are believed to be determined by genotype, *e.g.* eye color, hair color, height *etc.*

Introduction cont'd

- The only such dataset I found in the Internet was *HapMap* data (http://www.hapmap.org)
- This data was collected from people of different races and, unfortunately, have no phenotype traits attached to it, other than the race itself...
- So, for my project I determine the race of a person from his/her genotype

Disclaimer: I realize that this is a very sensitive subject, so I will be as politically correct as possible.

Human genome

- Human cells contain 23 pairs of chromosomes, with 22 pairs being almost identical and 23rd pair being sex chromosomes XX (females) or XY (males)
- Each chromosome contains a very densely packed DNA molecule which is, essentially, a very long chain (up to a few meters) of hundreds of millions amino-acids
- There are only 4 possible kinds of amino-acids in DNA: adenine, thymine, cytosine and guanine, abbreviated A, T, C, and G
- Thus, the DNA is simply 23 VERY LONG sequences of those letters, or rather couples of letters, since the chromosomes are coming in pairs

Example of a single data file:

rsid	NA06985	NA06991	NA06993	NA06994	NA07000	more people ...
rs10399749	CC	NN	CC	CC	CC	...
rs4030303	CC	CC	CC	CC	CC	...
rs940550	TT	TT	TT	TT	TT	...
rs13328714	AA	AA	AA	AA	AA	...
rs6683466	CC	CC	CT	CT	CC	...
rs12025928	GG	GG	GG	GG	GG	...
more rsids...	⋮	⋮	⋮	⋮	⋮	

- There are 23 chromosomes × 4 different groups of people = 92 files like this
- Some of the rsid's marked as non-readable ("NN") – I skipped them in analysis (cleaning data)
- Notice that the data are presented in "transposed" view – relatively few measurements (people) are given in columns and humongous amount of features (rsid's) – in rows. I believe this is a "standard" way of presenting genetic data.

Dataset description

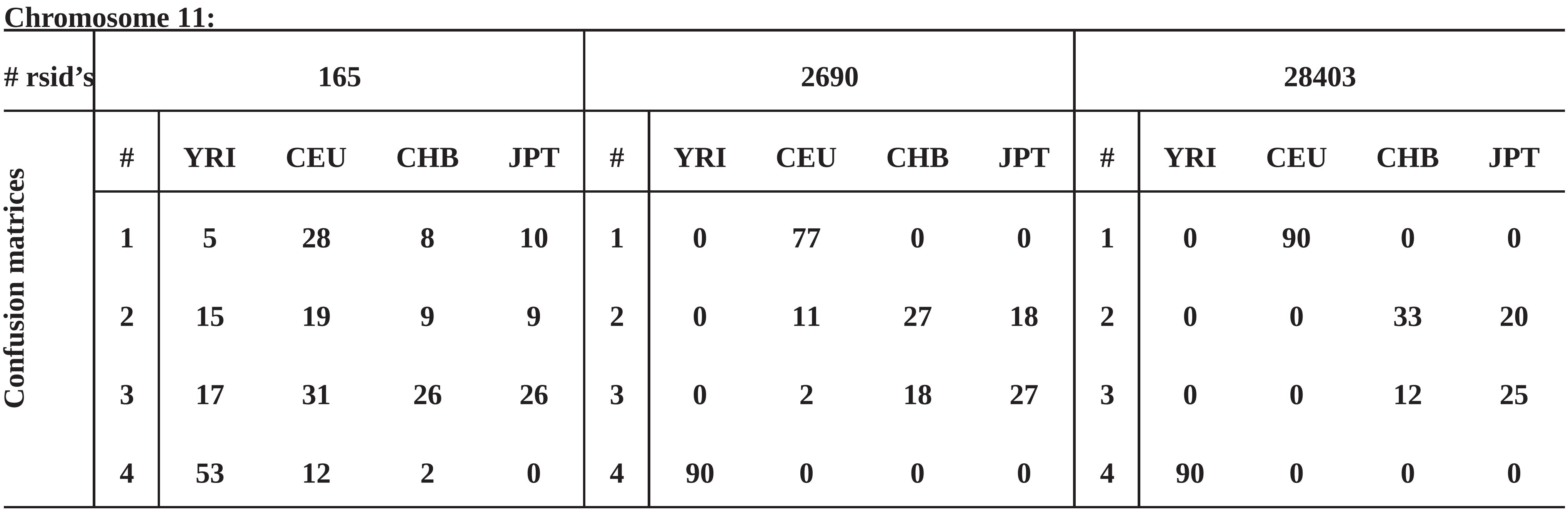
According to HapMap website, the DNA's in their dataset were obtained from the following groups:

- Yoruba in Ibadan, Nigeria (abbreviation: YRI) - 90 people
- Japanese in Tokyo, Japan (abbreviation: JPT) - 45 people
- Han Chinese in Beijing, China (abbreviation: CHB) - 45 people
- Utah residents with ancestry from northern and western Europe (abbreviation: CEU) - 90 people

Analysis

- I use two Data Mining techniques not sensitive to the (huge) number of features: KNN classification and K-means clustering
- Since all the “standard” packages in R and Python deal with numbers, I had to write my own code to deal with letters (main obstacle)
- I use the “hamming distance” between two DNA sequences which is determined as the total number of non-equal corresponding elements in those two sequences. This way the elements don’t have to be numbers.

K-means Clustering Results ($K = 4$)

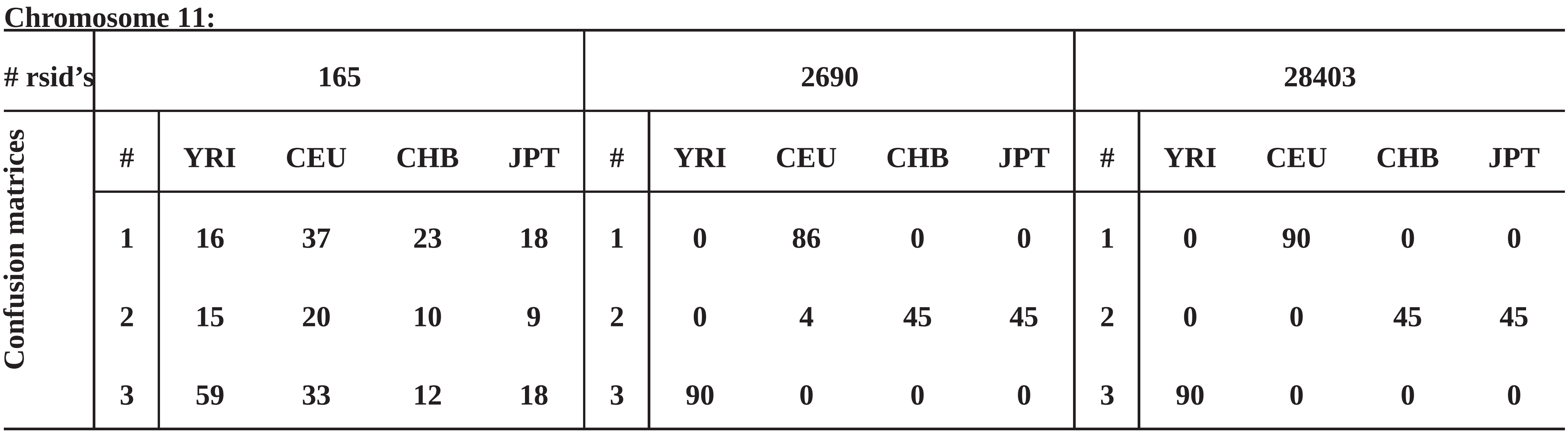


Chromosome 11:

# rsid's	165					2690					28403				
Confusion matrices	#	YRI	CEU	CHB	JPT	#	YRI	CEU	CHB	JPT	#	YRI	CEU	CHB	JPT
	1	5	28	8	10	1	0	77	0	0	1	0	90	0	0
	2	15	19	9	9	2	0	11	27	18	2	0	0	33	20
	3	17	31	26	26	3	0	2	18	27	3	0	0	12	25
	4	53	12	2	0	4	90	0	0	0	4	90	0	0	0

- Low number of used rsid's yields mixed confusion matrix
- Ten times higher number of used rsid's isolates YRI people and almost isolates CEU people
- Ten more times higher number of used rsid's completely separates YRI's and CEU's. The other two groups are still mixed ⇒ may be they cannot be separated?

K-means Clustering Results ($K = 3$)

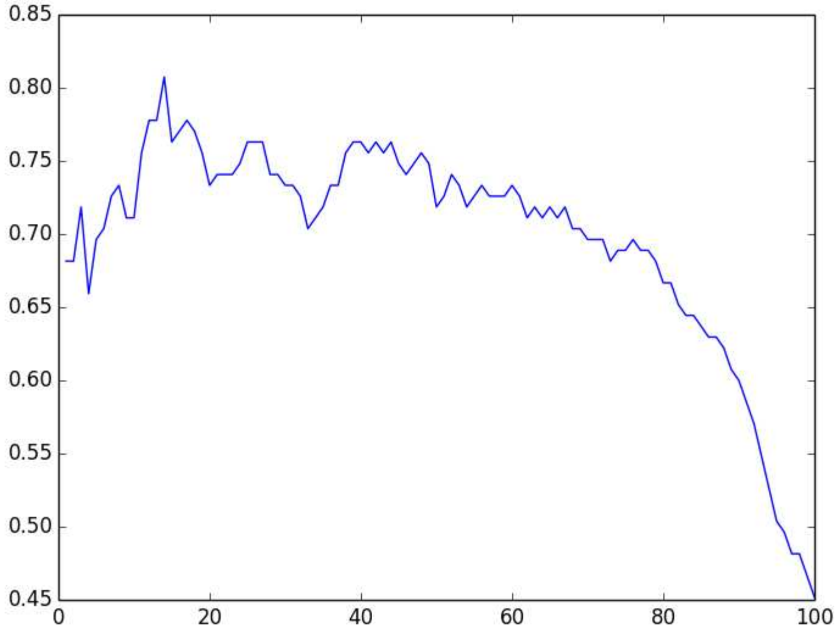
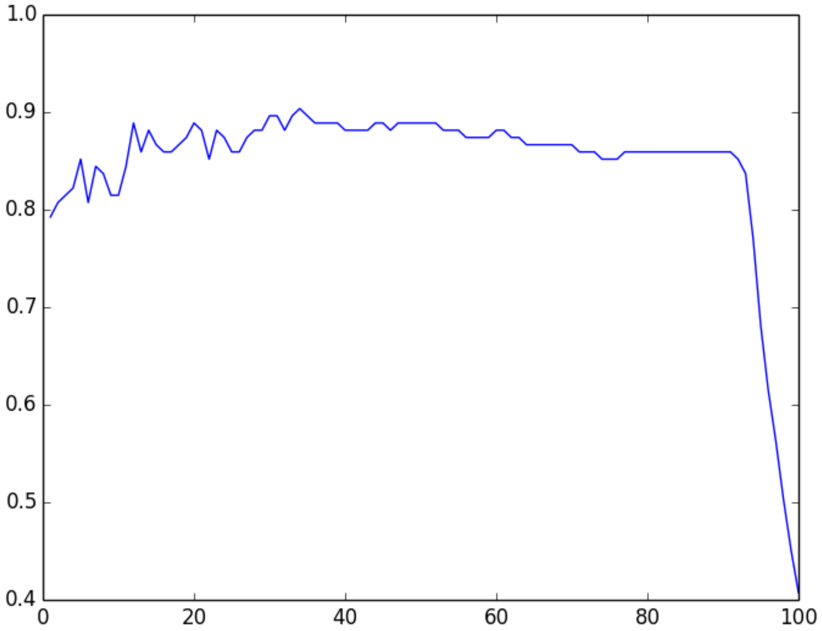


Chromosome 11:

# rsid's	165					2690					28403				
Confusion matrices	#	YRI	CEU	CHB	JPT	#	YRI	CEU	CHB	JPT	#	YRI	CEU	CHB	JPT
	1	16	37	23	18	1	0	86	0	0	1	0	90	0	0
	2	15	20	10	9	2	0	4	45	45	2	0	0	45	45
	3	59	33	12	18	3	90	0	0	0	3	90	0	0	0

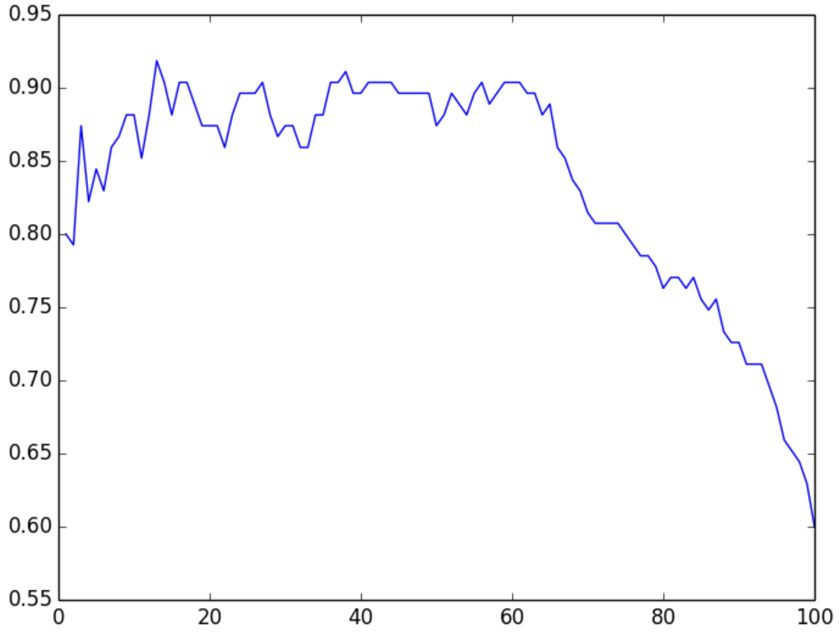
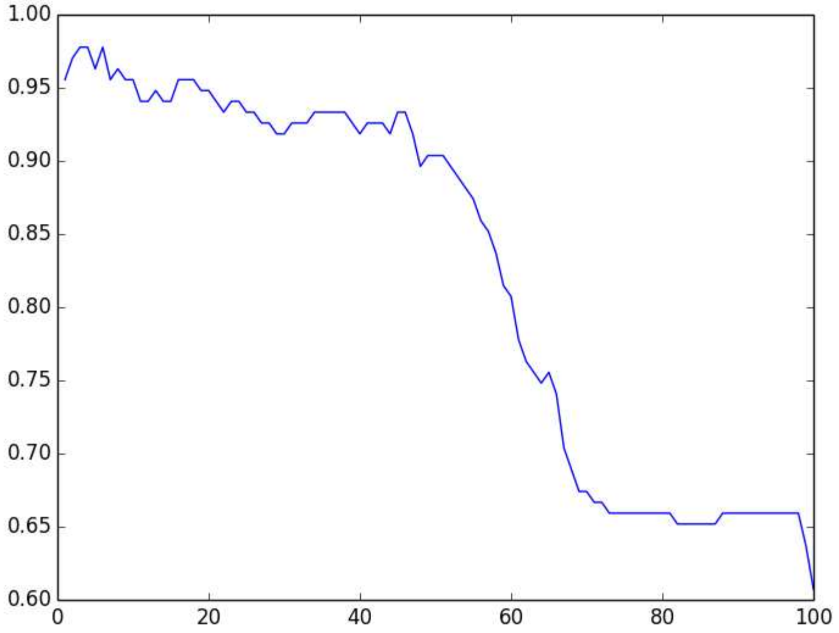
- Again, low number of used rsid's yields mixed confusion matrix
- Again, ten times higher number of used rsid's isolates YRI people and almost isolates CEU people
- And ten more times higher number of used rsid's completely separates YRI's, CEU's and "CHB + JPT" people.

KNN Classification Results for 4 Classes

#	165	2690
Accuracies vs. K		

- Low number of used rsid's yields 70%-80% prediction accuracy
- Ten times higher number of used rsid's raises prediction accuracy up to 80%-90%

KNN Classification Results for 3 Classes

#	165	2690
Accuracies vs. K		

- This time low number of used rsid's yields 80%-90% prediction accuracy
- Ten times higher number of used rsid's raises prediction accuracy up to 90%-97%

Conclusion

- The race of a person definitely can be derived from their genotype
- People from JPT and CHB groups seem to have genotypes belonging to the same cluster
- People from CEU and YRI groups have clearly distinguishable genotypes

Any questions?