Predictive Analytics to Determine the Potential Occurrence of Genetic Disease and their Correlation: Osteoporosis and Cardiovascular Disease

Kae Sawada
Computer Science Department
California State University, Los Angeles
Los Angeles, USA
ksawada@calstatela.edu

Michael W. Clark
Biology Department
Pasadena City College
Pasadena, USA
mclark7@pasadena.edu

Zilong Ye
Computer Science Department
California State University, Los Angeles
Los Angeles, USA
zye5@calstatela.ed

Nabil Alshurafa
Medicine and of Computer Science Department
Northwestern University
Evanston, USA
nabil@northwestern.edu

Mohammad Pourhomayoun
Computer Science Department
California State University, Los Angeles
Los Angeles, USA
mpourho@calstatela.edu

Abstract— **In this paper, a Predictive Analytics Model is designed, developed, and validated to determine the risk of manifesting osteoporosis in later life using big data processing. The proposed model leverages the novel genetic pleiotropic information in the 1,000 Genome Project of over 2,500 individuals world-wide. Also, the mutations associated with osteoporosis and cardiovascular disease are specifically analyzed. The study proposes the automatic histogram clustering as an effective and intuitive visualization method for high dimensional dataset. The results demonstrate a significant correlation between a person's regional background and the frequency of occurrence of the 35 Single Nucleotide Polymorphisms (SNPs) associated with osteoporosis and/or cardiovascular disease (CVD). Machine learning algorithms, such as Logistic Regression, Adaboost, and KNN are then applied to predict the occurrence of 7 osteoporosis-related-SNPs based on the existing CVD-related-SNPs input. Finally, the developed model is evaluated using a separate dataset obtained through Affymetrix microarray mRNA expression signal values for the specific SNP(s) in individuals with and without osteoporosis.**

Keywords- osteoporosis; Predictive Model; Genome Wide Association Study (GWAS); Clustering.

I. INTRODUCTION

This research is a continuation of an investigation previously presented in CSCI 2017 [34] by the same authors. The previous paper examined the dataset obtained from the 1000 Genome Project for its mutation variants with known relations to osteoporosis and cardiovascular disease. This paper proposes an novel approach to automatic data-driven clustering of histogram presented data for verification and validation of disease related expression in different human populations. As explained in later sections, high dimensional datasets can be effectively and intuitively visualized by the algorithm generating the histogram clustering. This automated process could aid in understanding existing correlations among various types of large datasets. Here, this automated clustering reproduced the manual clustering of the disease related SNPs geographic profiles. This study also confirmed the feasibility of the previously proposed predictor through a set of microarray analysis result dataset.

A. Osteoporosis and Space Exploration

Fractures, as a result of osteoporosis, have a significant negative impact on an individual's health, quality of life, and work performance. Some modern occupations inevitably expose workers to a significantly increased risk of developing osteoporosis. An obvious example is space flight. Within a few days of Zero Gravity (zero G), astronauts begin to lose both muscle and bone mass. There are also a zero G suppression effect on the immune system, as well as an increase in the aging process of particular cells. These effects occurred for all individuals who exposed themselves to zero G environment [25]. It is reasonable to assume that a predisposition to osteoporosis might increase the rate of occurrence of the condition. Having a reasonably reliable predictive model to reveal the predisposition for osteoporosis could be of great benefit for the future commercialization of space.

Over 7 TB of genomic data has been sequenced by the 1000 Genome Project [1]. These datasets consist of over 2,500 individuals from all around the world [2]. Such large

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-639-2

genomic datasets allow for the study of complex disease states, such as osteoporosis and cardiovascular disease. These pathologies involve not only multiple gene mutation interactions, but also nutritional and age-related components. [9][20].

There are over 100 genes that are proven to be related to osteoporosis, according to Genome-Wide Association Studies (GWAS) [7]. However, many of these gene mutations still maintain some level of phenotypic ambiguity [22]. Understanding and treating such complex conditions could benefit through modern Machine Learning analysis. An effective method for early forecast of a patient-specific frequency of occurrence for a particular genetic disorder in later life would allow for appropriate preventative measures to be followed.

B. Mutations: SNPs

This paper analyzes 35 SNPs, that are commonly observed as genetic disease related mutations. It is a mutation of one base for another, which occurs in more than one percent of the general population [14]. 7 of these SNPs have direct indications in the expression of osteoporosis [9], while the other 28 SNPs have implications in both CVD and osteoporosis [4].

C. Preceding Related Work

i. Osteoporosis-related-SNP selection

The 7 osteoporosis-related-SNPs were chosen based on the study led by Hsu et al [9], published in 2010. The phenotypic association of these SNPs to osteoporosis was demonstrated by the GWAS study [26].

ii. Genetic Pleiotropy

Using False Discovery Rate (FDR) statistical methods, Reppe et al [4] revealed a potential genetic link between Cardio-Vascular Disease (CVD) and Osteoporotic conditions. In this paper, the potential mutant gene interactions between the osteoporosis-related SNPs and CVD SNPs are analyzed with: big data processing and analytics, predictive analytics based on machine learning algorithms, data visualization, and clustering.

This paper is organized as follows: in Section II, the methods and materials employed during the analysis and experimental developments are introduced. Included topics are the datasets and methods used, feature selections and dataset label, and evaluation methods. In Section III, the results and observations are discussed.

II. METHODS AND MATERIALS

This section discusses the datasets employed, the design of the analysis and research, algorithms and techniques utilized in the study so far.

A. Datasets

Two big datasets have been used in this paper:

The first one is the genotypes from the 1000 Genome Project for the 35 SNPs. This dataset includes 2504 human subjects, containing both male and female from 26 regions worldwide [2].

The second dataset is from Reppe et al's study mentioned in Section I-C-ii [28]. The samples of this dataset are collected from the 84 post-menopausal females between the ages of 50 to 86 years old in Lovisenberg Diakonale Hospital located in Sweden. There are two components to this dataset:

(1) The result of Affymetrix Microarray analysis of the patients, the Affymetrix microarray signal values per sample, one ".CELL" file per patient.

(2) A set of biopsy results, sample ID (anonymous), age, gender, and the biopsy results of bone density scores, both T- and Z-scores, consisting of average neck, total hip, and average spine of each subject.

In this paper, all ".CELL" files from (1) were processed and interfaced with the library, pd.hg.u133.plus.2 [27], to obtain the gene symbol per an array cell. Then, it was interfaced with SNP identifier (e.g., #rsxxx..x, with x composing an integer) to acquire the existing SNP(s) data per each sample. In acquisition of SNP existence, the signal threshold value was determined the average score of all samples per column (per array cell). Such determination was made based on the preceding study which defined signal thresholds in DNA microarray analysis [29].

The label per sample was acquired based on the data from (2). Samples with the T-scores of -2.5 or less in one or more of neck, hip, and spine are marked as osteoporotic. The threshold was determined, following the World Health Organization (WHO) international reference standard for osteoporosis diagnosis [30]. T-scores were used instead of Z-scores, based on the sample's age and the guidelines provided by WHO.

B. Method Design

i. Problem Definition

The previous study investigated the occurrence of 7 selected osteoporosis-related SNPs [3] and its correlation to 28 CVD with 6 subdivisions, related SNPs [4], shown in Table 1 and 2. The CVD related SNPs were divided into six subcategories: High Density Lipids proteins cholesterol (HDL), Low Density Lips proteins cholesterol (LDL), Systolic Blood Pressure (SBP), Diatonic Blood Pressure (DBP), Type 1 Diabetes (T1D), and Triglycerides (TG). The table of osteoporosis-related SNPs describes the SNP identification number, normal (ancestral) base, high-risk (mutated) base, and the homozygous base pairs that are associated with a high risk of Bone Mineral Density (BMD) loss, and consequently the development of osteoporosis. (See Section I-C-i for osteoporosis-related SNPs). In this paper, we also validate the predictor design by utilizing the new dataset introduced in Section I-A.

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-639-2

ii. Feature Selection and Label

As mentioned in the Section II-C-ii, 28 CVD-related SNPs, and age of each subject were fed to the predictive model. The output of this predictor is a boolean value, whether the individual has developed osteoporosis or not.

iii. Predictive Analytics Algorithms

To develop a software method for predicting a tendency for disease occurrence, various machine learning algorithms were tried for building the predictive model: KNN (the best value of K determined by trial and error was K=9), Logistic Regression, Decision Tree, Naive Bayes, Adaboost, Random Forest, and Support Vector Machine (the best parameters of SVM determined by Grid Search include C=1, kernel='rbf', gamma=0.0007). A systematic aggregation of these classifier results is future work, for example using an ensemble learning algorithm.

TABLE 1: SNP ASSOCIATED WITH OSTEOPOROSIS

SNP (rs ID)	Ancestral allele	Mutated allel	Possible pair	High Risk Genotype	Phenotype (associated condition)
rs2278729	G	A	AA, GG, AG, GA	AA	Osteoporosis
rs12808199	A	G	AA, GG, AG, GA	GG	Osteoporosis
rs7227401	G	T	GG, TT, GT, TG	TT	Osteoporosis
rs494453	T	C	TT, CC, TC, CT	CC	Osteoporosis
rs12151790	G	A	AA, GG, AG, GA	AA	Osteoporosis
rs2062375	C	G	CC, GG, CG, GC	GG	Osteoporosis
rs17184557	T	A	TT, AA, TA, AT	AA	Osteoporosis

TABLE 2: SNP ASSOCIATED WITH CVD

SNP (rs ID)	Ancestral allele	Mutated allel	Possible pair	High Risk Genotype	Phenotype (associated condition)
rs4957742	A	G	AA, GG, AG, GA	GG	DBP
rs665556	C	T	TT, CC, TC, CT	TT	DBP
rs10779702	A	G	AA, GG, AG, GA	GG	HDL
rs12137389	T	C	TT, CC, TC, CT	CC	HDL
rs9309664	G	A	AA, GG, AG, GA	AA	HDL
rs7594560	T	C	TT, CC, TC, CT	CC	HDL
rs10953178	C	T	TT, CC, TC, CT	TT	HDL
rs980299	T	C	TT, CC, TC, CT	CC	HDL
rs10746070	T	C	TT, CC, TC, CT	CC	HDL
rs7175531	C	T	TT, CC, TC, CT	TT	HDL
rs3198697	C	T	TT, CC, TC, CT	TT	HDL
rs756632	C	T	TT, CC, TC, CT	TT	HDL
rs4820539	G	A	AA, GG, AG, GA	AA	HDL
rs6583337	G	A	AA, GG, AG, GA	AA	LDL
rs11809524	C	T	TT, CC, TC, CT	TT	SBP
rs11675051	G	A	AA, GG, AG, GA	AA	SBP
rs13005335	A	G	AA, GG, AG, GA	GG	SBP
rs12995369	A	G	AA, GG, AG, GA	GG	SBP
rs10464592	G	A	AA, GG, AG, GA	AA	SBP
rs1670346	A	G	AA, GG, AG, GA	GG	SBP
rs13272568	A	C	AA, CC, AC, CA	CC	SBP
rs600231	G	A	AA, GG, AG, GA	AA	SBP
rs258415	C	A	AA, CC, AC, CA	AA	SBP
rs11614913	C	T	TT, CC, TC, CT	TT	SBP
rs199529	C	A	AA, CC, AC, CA	AA	SBP
rs8090312	G	A	AA, GG, AG, GA	AA	T1D
rs2282930	G	A	AA, GG, AG, GA	AA	TG
rs10851498	T	C	TT, CC, TC, CT	CC	TG

Affymetrix microarray consists of a grid of oligonucleotide probes produced to have a known DNA sequence. The grid Microarray thus holds a specific SNP mutation at a specific locus on the grid. Preparations of labelled mRNA (cDNA/cRNA) taken from the individual patients can then be exposed to the entire grid containing the variety of SNP mutations. Identification of a specific SNP in the patient is determined by the measured level of hybridization with the corresponding target grid position and the labelled cDNA/cRNA. The corresponding SNP IDs were mapped through the affy ID and a gene symbol that are assigned to each cell, as well as the manual mappings of the target SNPs through a capability of the genome browser provided by University of California, Santa Cruz [31].

iv. Evaluation Method

The accuracy of the predictive model was evaluated using 30-fold cross-validation [32]. The performance was measured by the Area Under Curve (AUC) of the Receiver Operator Curve (ROC) . ROC AUC method accounts for the number of true positive predictions against the number of false positive predictions. Statistically, only about 30% of all postmenopausal women develop osteoporosis during their life time [35] thus, the testing dataset is unbalanced. The ROC-AUC can be an effective metric to measure and present the performance of the prediction model.

v. Phenotype Expression Measurement

The predictor execution on this particular dataset was expected to evaluate the geno-pheno-transfer rate. This will also validate the predictor developed in the previous study.

vi. Visualization

This study visualizes the dataset of over 30 essential features by generating various histograms and applying a K-mean clustering algorithm to automatically cluster the resulting plots into groups in a data-driven manner. As demonstrated in the result section, existing correlations among the dataset is clearly displayed through this visualization method. The proposed visualization method can aid the observers in developing an intuitive understanding of the dataset, and effectively mirroring the datasets' characteristics and patterns in them.

III. RESULTS AND DISCUSSION

This section discusses the result obtained from the predictor model and the histogram clustering automation.

A. Predictor results – Dataset from 1000 Genome Project

In the previous study, the predictive models were built and tested on 32 sets of inputs, where each set is an element of the powerset of 6 CVD related conditions. The results obtained from various combinations of SNP inputs showed a strong correlation between the 7 osteoporosis-related SNPs and the HDL2 SNPs. Similarly, another strong correlation was found between the 7 osteoporosis-related SNPs and SBP2 SNPs. The best classification result was achieved

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-639-2

using Logistic Regression classifier with the accuracy of 0.7769. Furthermore, there is a strong implication that the likelihood of developing osteoporosis could be determined/predicted by the existing CVD-related factors such as Cholesterol levels, Blood Pressure, and Triglycerides, levels [34].

B. Predictor results – Affy dataset

The performances of the predictive model for 3 different scenarios are listed in the following. The performance measurement was obtained through ROC AUC as mentioned in Section II-B-v. Unlike the previous experiment, this experiment setting did not distinguish high-risk heterozygous pair vs. high-risk homozygous pair. Here are the results for 3 different scenarios:

Scenario 1: Osteoporosis SNPs only – AUC = 0.7285
Scenario 2: CVD SNPs only – AUC = 0.7569
Scenario 3: Both CVD and Osteo SNPs - AUC = 0.8571

The predictor strategy developed here has been confirmed with the available disease related database of Reppe et al. The predictor was shown to be correct in the majority of the patients. These results also confirm the correlation of osteoporosis SNPs with CVD SNPs reported in the above Section. In most cases, the chance of a mutant genotype expressing its aberrant phenotype is much less than 100%. As a well-studied example, the two mutations related to Breast Cancer, BRCA – 1 and BRAC – 2, show such facts. The chances these two mutations expressing their phenotypes by the time a woman is 70 years old are as follows, according to a report by the Susan G. Komen Foundation [33]:

- BRCA 1 - 55%
- BRCA 2 - 45%

The accuracy of the predictor developed with this paper of close to 70% demonstrated an effective prediction. Such a score would thus be reasonably sufficient to caution individuals about having a higher risk of BMD loss. With such knowledge, individuals can take the necessary preventative measures to prevent the development of undesirable conditions and disease, starting at an early stage of their lives.

C. Data Visualization

Various histograms were drawn and clustered effectively to demonstrate the correlations among these high dimensional datasets. The results demonstrated distinct region-specific SNP frequency profiles. They confirmed the genetic links between osteoporosis, and displayed the intricacy of the interrelationship among the SNPs. Grouping the results demonstrated the divides in sample's SNP characteristics. These results are consistent with the idea that an individual's genetic and gene variant profile is related to the region in which his/her ancestors came from. As it was introduced in the previous study [34]. Figure 1 demonstrates the claimed regional divides clearly. CVD (Systolic Blood Pressure)-related SNPs are colored in green, and osteoporosis-related SNPs are in orange. The trends that are obvious in all 28 regions are similar here. Again, the regional divides are well maintained. This type of histogram comparison table was originally introduced in the previous study [34].

The result was also consistent with the fact that European women show higher incidents of osteoporosis and hip fractures when compared with populations from other regions. Europeans have higher frequency of osteoporotic mutations than African [8].

D. Clustering Automation

An automated clustering process was developed for grouping the histograms to better visualize the results. The results of clustering clearly confirms the regional divide of the histogram profiles. Figure 2 displays a sample result of this automated clustering. Sample groups obtained in automated histogram clustering result. The results verify the regional divide of osteoporosis- and CVD-related SNP profiles. The left most column of Figure 2 shows European region, Iberian region in Spain, Utah region (CEPH) with Northern and Western European Ancestry, British region in England and Scotland, Finnish region in Finland, Toscani region in Italia. The middle column shows Bengali region of Bangladesh, Sri Lankan region, Tamil from the UK, Indian region, Telugu region in the UK, Gujarati Indian ancestry in Houston, Texas, Punjabi region from Lahore, Pakistan. The right most column shows Peruvians from Lima, Peru, Puerto Ricans from Puerto Rico, Mexican Ancestry from Los Angeles USA, Colombians from Medellin, Colombia.

To demonstrate the value of this clustering automation, the final number of clusters to be formed by the algorithm was varied. As shown in Figure 3, when the number of clusters is set to 12, the algorithm groups South Asian regional SNP profiles and American regional SNP profiles into two separate groups, perfectly distinguishing the two separate regions. However, when the number of cluster is set to 11,

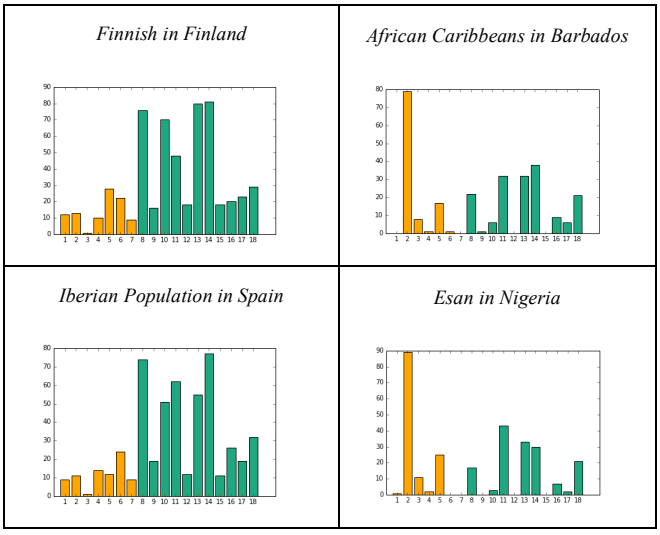

Figure 1: SBP SNP Profile, European Region vs. African Region

Copyright (c) IARIA, 2018. ISBN: 978-1-61208-639-2

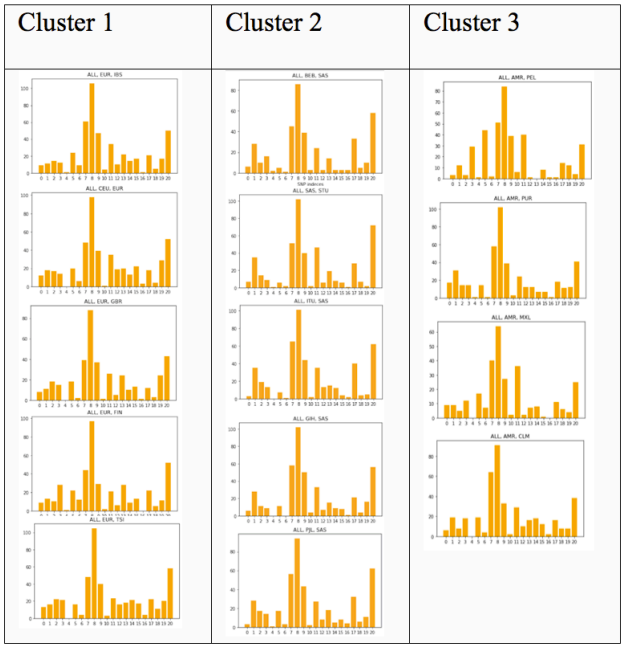

Figure 2: A Sample Histograms Demonstrating Regional Devides

the South Asian regional profiles and American regional profiles are bundled up into a single group. Such a result shows the possibility of further dissection of the regionalism seen here by the automated process. Automation of the histogram profiles also allowed increased number of SNPs to be analyzed all at once.

E. Summary of the Results

i. The high correlation between a person's regional background and the occurrence of the selected 35 SNPs associated with osteoporosis and/or CVD shown in the previous investigation were confirmed by this new methodology. It was confirmed that the highest fracture rates and osteoporosis related SNPs are found in white women of European descent; African Americans tend to have higher Bone Mineral Density (BMD) with few osteoporosis related SNPs [8], as introduced in the previous study [34].

ii. The set of CVD-/osteosis-related SNPs from Reppe et al's dataset can be used to predict a person's likelihood of developing osteoporosis with the accuracy of approximately 64% on their own.

iii. The CVD-osteoporosis related SNPs from Reppe et al's dataset outperformed osteoporosis-related SNPs (see Section I-C-i) in predicting osteoporosis.

IV. Conclusions and Discussions

This paper extended on the previous publication [34] with the following contributions: 1) Validated the predictor model introduced in the previous publication based on the second dataset, using "Area Under Curve" of the "Receiver Operator Curve" evaluation method. The model predicts an individual's likelihood of BMD loss over time. 2) Confirmed that an individual's predisposition to osteoporosis can be linked a predisposition to CVD related conditions. 3) Confirmed a correlation between individual's regional background and the SNP occurrence profiles.

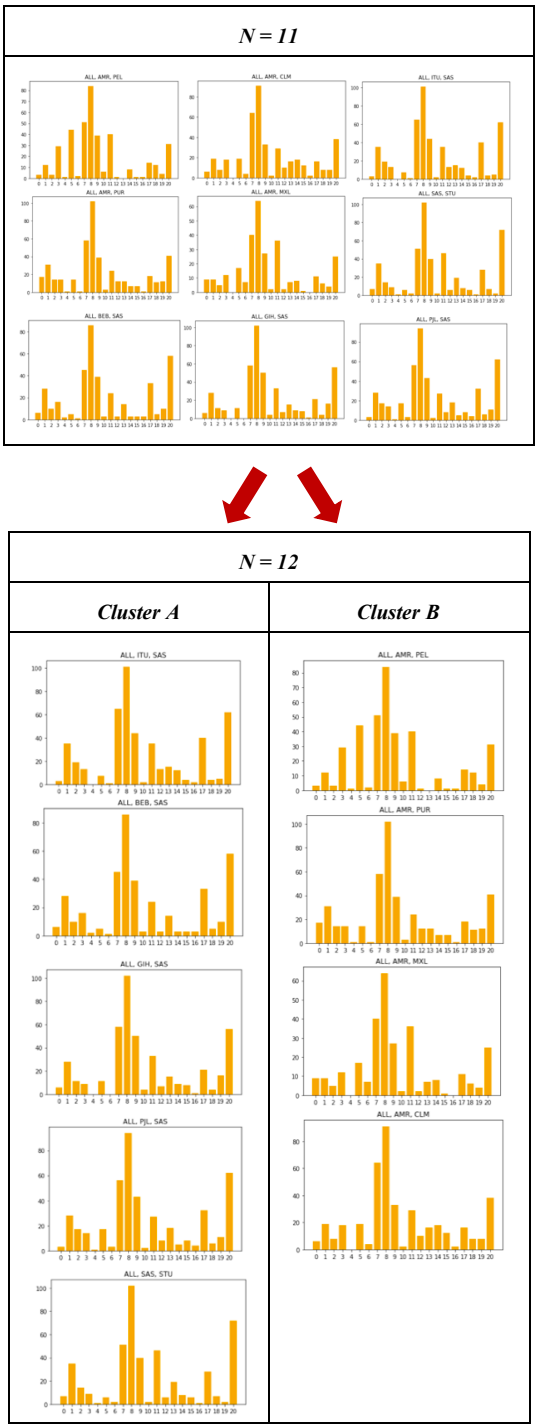

Figure 3: A Sample Auto-Clustering Result

4) Automated clustering of the histograms of SNP occurrence profiles, using the K-mean clustering algorithm. This study also proposed a histogram auto-clustering

approach to effectively visualize the correlations that exists among a high dimensional dataset.

The United Kingdom is now half way through its 100,000 genome project. The goal of obtaining this massive human genome dataset was to determine genetic predisposition to genetic related diseases, such as cancer, CVD and osteoporosis. The automated systems developed here will be of great assistance in achieving these goals.

Revisiting our future in space, as introduced in Section I-A, the exact effects of the outer space specific environments on human biological systems are unknown. Understanding such effects on skeletal system, nervous system, reproductive system, and our genome is crucial when attempting to adapt to an unknown environment. NASA and other space agencies are actively investigating these issues. Being able to use an astronaut's genetic predisposition to predict how that individual's body might respond to zero or low Gravity could help individuals take proper precautions, improve the outcome of space flight, and support the expansion of humanity into space. It will also provide an equal and safer opportunity to everyone who wishes to go to outer space. As we explore into various unknown environments of outer space, capabilities to quickly collect and analyze large amounts of information will be essential. Placing such data into the learning algorithms to update the predictor will allow the explorer appropriately respond to the new environments. With such a system on board, astronauts' well-being and success of the mission can be supported.

REFERENCES

[1] IGSR and the 1000 Genomes Project. [Online] Available at: http://n.internationalgenome.org/ [retrieved: Nov. 2017].

[2] Which populations are part of your study? [Online] Available at: http://www.internationalgenome.org/faq/which-populations-are-part-your-study [retrieved: Nov. 2017].

[3] D. Karasik, H. Yi‐Hsiang, Y. Zhou, L. A. Cupples, D. P. Kiel, and S. Demissie. "Genome‐wide pleiotropy of osteoporosis‐related phenotypes: the Framingham study." Journal of Bone and Mineral Research 25, no. 7, pp. 1555-1563, 2010.

[4] S. Reppe, Y. Wang, W.K. Thompson, L. K. McEvoy, A. J. Schork, V. Zuber et al. GEFOS Consortium. "Genetic sharing with cardiovascular disease risk factors and diabetes reveals novel bone mineral density loci." PloS one 10, no. 12, e0144531, 2015.

[5] O. Hitomi, S. Sasaki, H. Horiguchi, E. Oguma, K. Miyamoto, Y. Hosoi, M. K. Kim, and F. Kayama. "Dietary patterns associated with bone mineral density in premenopausal Japanese farmwomen–." The American journal of clinical nutrition 83, no. 5, pp. 1185-1192, 2006.

[6] J. B. Richards, H. F. Zheng, and T. D. Spector, "Genetics of osteoporosis from genome-wide association studies: advances and challenges," Nature Reviews Genetics, 2012.

[7] GWAS Catalog. [Online] Available at: https://www.ebi.ac.uk/gwas/ [retrieved: Nov. 2017].

[8] N. Hae-Sung, S. S. Kweon, J. S. Choi, J. M. Zmuda, P. C. Leung, L. Y. Lui, et al. "Racial/ethnic differences in bone mineral density among older women." Journal of bone and mineral metabolism 31, 2013.

[9] H. Yi-Hsiang, et al. "An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits." PLoS genetics 6, no. 6, e1000977, 2010.

[10] L. Langsetmo, et al. "Using the same bone density reference database for men and women provides a simpler estimation of fracture risk." Journal of Bone and Mineral Research, no.10, pp. 2108-2114, 2010.

[12] J. Greenbaum, K. Wu, L. Zhang, H. Shen, J. Zhang, and H. W. Deng. "Increased detection of genetic loci associated with risk predictors of osteoporotic fracture using a pleiotropic cFDR method." Bone 99, pp. 62-68, 2017.

[13] D. Karasik, and M. Cohen-Zinder. "Osteoporosis genetics: year 2011 in review." BoneKEy reports 1, no. 8, 2012.

[14] U.S. National Library of Medicine | What are Single Nucleotide Polymorphisms (SNPs)? [Online] Available at: https://ghr.nlm.nih.gov/primer/genomicresearch/snp [retrieved: 12 Nov. 2017].

[15] Y. H. Hsu, et al. "An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits." PLoS genetics 6, no. 6, e1000977, 2010.

[17] I. M. Meyer, and R. Durbin. "Gene structure conservation aids similarity based gene prediction." Nucleic acids research 32, no. 2, pp. 776-783, 2004.

[18] R. J. Carter I. Dubchak, and S. R. Holbrook. "A computational approach to identify genes for functional RNAs in genomic sequences." Nucleic acids research 29, no. 19, pp. 3928-3938, 2001.

[19] T. Chen, M. Y. Kao, M. Tepel, J. Rush, and G. M. Church. "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry." Journal of Computational Biology 8, no. 3, pp. 325-337, 2001.

[20] US Department of Health and Human Services. "Bone health and osteoporosis: a report of the Surgeon General." Rockville, MD: US Department of Health and Human Services, Office of the Surgeon General 87, 2004.

[21] P. Salari, and M. Abdollahi. "The influence of pregnancy and lactation on maternal bone health: a systematic review." Journal of family & reproductive health 8, no. 4, p. 135, 2014.

[22] L. Qin, Y. Liu, et al "Computational characterization of osteoporosis associated SNPs and genes identified by genome-wide association studies." PloS one 11, no. 3, e0150070, 2016.

[23] E. R. Spector, S. M. Smith, and J. D. Sibonga. "Skeletal effects of long-duration head-down bed rest." Aviation, space, and environmental medicine 80, no. 5, A23-A28, 2009.

[24] The Matplotlib API [Online] Available at: https://matplotlib.org/2.0.2/api/index.html [retrieved: Nov. 2017].

[25] E. Roberge. "The Gravity of It All: From osteoporosis to immunosuppression, exploring disease in a microgravity environment holds promise for better treatments on Earth." IEEE pulse 5, 2014.

[26] Ensemble Genome Browser [Online] Available at: https://uswest.ensembl.org/index.html [retrieved: Jan. 2018].

[27] Bioconductor 3.6 Annotation package, pd.hg.u133.plus.2 [Online] Available at: http://bioconductor.org/packages/release/data/annotation/html/pd.hg.u133.plus.2.html [retrieved: Jan. 2018].

[28] E-MEXP-1618 - Transcription profiling of bone biopsies from postmenopausal females identifies 8 genes highly associated with bone mineral density [Online] Available at: https://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-1618/files/ [retrieved: Jan. 2018].

[29] M. Bilban, L. K. Buehler, S. Head, G. Desoye, and V. Quaranta. "Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer." BMC genomics 3, no. 1, 19, 2002.

[30] 2015 ISCD Official Positions – Adult [Online] Available at: https://www.iscd.org/official-positions/2015-iscd-official-positions-adult/ [retrieved: Jan. 2018].

[31] UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly [Online] Available at: https://genome.ucsc.edu/ [retrieved: Jan. 2018].

[32] The documentation for *sklearn.model_selection.cross_val_score* [Online] Available at: http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html [retrieved: Jan. 2018].

[33] Susan G. Komen Available at: ww5.komen.org [retrieved: Jan. 2018].

[34] K. Sawada, M. W. Clark, N. Alshurafa, and M. Pourhomayoun, "Analyzing the Mutation Frequencies and Correlation of Genetic Diseases in Worldwide Populations Using Big Data Processing, Clustering, and Predictive Analytics," International Conference on Computational Science and Computational Intelligence, pp. 1459 – 1464, 2017.

[35] Epidemiology by International Osteoporosis Foundation [Online] Available at: https://www.iofbonehealth.org/epidemiology.