Extraction of information and facts from data mining of random sequences for undergraduate research

Dr. sunil Dehipawala, Queensborough Community College

Sunil Dehipawala received his B.S. degree from University of Peradeniya in Sri Lanka and Ph.D from City University of New York. Currently, he is working as a faculty member at Queensborough Community College of CUNY.

Dr. Raul Armendariz, Queensborough Community College

Assistant professor of physics

Mr. George Tremberger Jr, CUNY Queensborough Community College
Prof. Tak Cheung, CUNY Queensborough COmmunity College

Tak Cheung, Ph.D., professor of physics, teaches in CUNY Queensborough Community College. He also conducts research and mentors student research projects.

Extraction of information and facts from data mining of random sequences for undergraduate research

Sunil Dehipawala, Raul Armendariz, George Tremberger, David Lieberman, and Tak Cheung

CUNY Queensborough Community College Physics Department

Abstract

A general method to extract information and facts from data mining of random sequences in biology and astronomy has been developed. The random sequence analysis has been implemented in several NSF-REU projects using NIH and NASA databases. Examples of RNA sequence with reference to Shannon-entropy based bioinformatics and SDO magnetic topology analysis with reference to solar physics are presented. The contrast to Brookhaven Synchrotron and high energy physics data analysis is also discussed. The feedback of administrating REU projects to our physics teaching for engineering students was found to be valuable and the examples of magnetic reconnection and geomagnetic induced current are presented. Therefore community college REU project provides connectedness awareness in the linking of previous published reports, critical thinking in result interpretation, and career development when going onto a senior college REU program, the top three benefits of college education, according to a 2016 July Money Magazine "Value of College" survey.

Keywords

Random sequence analysis, mRNA sequence data, NASA SDO data, NR2F2 gene.

Introduction

The mixing of noise and signal is a fact in science and whether it is possible to extract useful information from the noise component is an issue of particular interest. A well-known 2016 example is in the detection of gravitational wave with substantial noise [1]. The recent discovery of a 4-exoplanet system in tau-Ceti is another example [2]. Our community college has a NSF- REU program that offers research opportunities for community college engineering and science students in the New York Tristate area. Among our various REU projects, the extraction of information and facts from data mining of random sequences in biology and astronomy is discussed. In particular the NASA Solar Dynamics Observatory SDO and NIH Genbank RNA data are used as illustration.

NASA Solar Dynamics Observatory SDO data analysis

NASA has been posing solar eruption images in its Outreach program that attract students' attention. A student project would examine the data fluctuation after subtracting out the

decreasing/increasing trend, hence a random series analysis, a novel approach in community college student research project as far as we know. In the area of stock data, taking away the trend could reveal correlated non-random fluctuation known as volatility in which further analysis could suggest some underlying nonrandom issues. The Black Sholes model used in option trading, related to the 1997 Economics Nobel Prize, contains a distribution that can be described by a differential equation. A similar approach for the analysis of SDO images has been reported by us earlier [3]. The adaptation to community college student research was done by calibrating original data using NASA packages, such as The Interactive Data Language (IDL) and/or The Interactive FITS File Editor (Fv), with the posed RGB image of an astronomical object so that popular image processing software such as ImageJ could be used [4, 5, 6].

The SDO data can be used to study magnetic topology via the formula of quasi- separatrix layers QSL or quashing factor Q. The grouping of field lines into separate bundles that connect disparate regions on the solar surface as measured by SDO has been performed [7, 8]. In operation the squashing factor involves the measurement or computation of the distance between two closely-spaced field lines with their conjugate footpoints. The Pasco equipotential field line experiment has been one of our standard labs with simple contour line tracing, and students would have no difficulty in locating the contour lines in a SDO dataset. Furthermore the field line random walk property study in community college REU projects can be performed with a numerical correlation approach on Excel-VBA and/or Matlab performs based on published literature [9]. The galaxy magnetic topology from light polarization measurement in gravitational lensing setting can be studied when the field lines are already published [10], and simplification for extension to high school project in a K-12 education scheme could be performed using Science News contents such as those galaxy magnetic field examples posted on Phys.org [11, 12].

NIH mRNA sequence analysis

The analysis of mRNA sequence could shed light on the non-coding region which is important in control and gene regulation. The A, T, C, G nucleotide sequences for homolog sequences have been examined for the difference in the non-coding regions because the homolog sequences share homologous protein structure and NIH confirms the laboratory verifications and routinely updates the mRNA database. Operationally speaking, when two random variables form a regression with R-sq > 0.9, the issue of what underlies the high correlation arises. The Shannon entropy values (p * log p with p being the probability) of the mononucleotide and the di-nucleotide of a mRNA sequence can be computed. The mononucleotide entropy has 2 bits per nucleotide maximum value while the di-nucleotide entropy has 4-bits per pair maximum value given the 16 possibility AT, AC, AG, etc.

It was reported on August 2017 that female mouse embryos would actively remove male reproductive systems using a NR2F2 gene, rewriting the textbook knowledge of using androgen as a default since 1950 [13]. A study of mRNA sequence on the species sharing the homolog protein structure was conducted and our novel method is described below. The NR2F2 homolog information is listed on NIH Genbank webpage with the web address given as "ncbi.nlm.nih.gov/homologene/7628". The mRNA mononucleotide entropy was found to have a high correlation with the dinucleotide entropy for the five studied sequences, as shown in Figure 1. The mononucleotide entropy is related to point mutation and/or SNP while the di-nucleotide entropy value is related to epigenetics through methylation [14]. When two random variables are

correlated as experimental fact, there could be an underlying structure governing the non-coding region. In contrast, the correlation of mononucleotide entropy with sequence length was founded to be low at R-sq = 0.23.

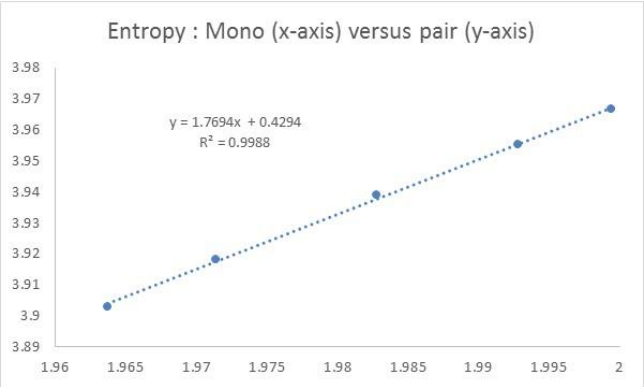

Figure 1: Mononucleotide entropy (x-axis) versus di-nucleotide entropy (y-axis) of NR2F2 mRNA sequences associated with the homolog proteins for human, rat, chicken, zebrafish and fruit fly.

Discussion

Our REU program reviews the available projects and statistical methods to the recruited students in the first week. Students are then asked to list their project choices at the end of the first week after discussion sessions with the related faculty supervisors. Our previous REU program data showed that students usually would be worrisome about selecting information extraction projects as their first priorities. The reasons could include a bias attitude towards data mining, difficulty in the perceived computation, etc. Our current REU program introductory week now includes the explanation of information extraction in terms of the spreadsheet formulas and Visual Basic Application VBA engine in Microsoft Excel.

The REU students are required to attend a weekly meeting and discuss their projects in PowerPoint presentations. The information extraction presentations illuminate a deeper understanding of signal processing among all REU students. One may argue that the information extraction projects are too abstract to benefit community college students in comparison to other projects with wet-laboratory objectives. The students would rely totally on the faculty supervisors for mathematical facts on random series, resulting in less independent thinking development. However it has been reported in 2014 that "American universities' focus on independence undermines the academic performance of first-generation college students.[15]" Recently Vox.com has a summary article saying that the university administrator focus on independence is a major contributing factor that college education has become an inheritance, which first generation students do not have [16, 17].

The offering of information extraction projects is consistent with the following two facts. Firstly, the Vox.com survey result showed that first generation students prefer "helping family after graduation" as the first reason for attending college when compared to the continuing generation students. Secondly, the building of a biomedical data science workforce is favored by NIH [18]. Indeed the extraction methodology presented above aims to encourage instructors to include information extraction as high impact projects for first generation students.

The research of our students could be viewed as a "college-parallel" to the graduate school research in the astronomy examples, and as a novel investigation of mRNA Shannon entropy based bioinformatics. Each student project presentation includes the demonstration of the relationship between the modeled parameters with simple calculations at introductory physics level so as to reinforce the education component beyond numerical model fitting.

The contrast to Brookhaven Synchrotron and high energy physics data analysis in the other REU projects is informative in terms of the college courses taken by a student. Synchrotron data analysis would proceed with standard software packages with less demand on programming skill. The Extended X-Ray Absorption Fine Structure (EXAFS) noise smoothing procedure is under software control and the Fourier transform from wave vector space to bond length space could be performed without having a math course on periodic functions. The high energy physics data analysis requires ROOT special programming language that demands strong programming skill. Regardless of which particular project is being selected by a student, each student would need to know how to use Excel to calculate statistics such as p-value. One of the free online literature on Excel statistics has been used [19].

The feedback of REU to our teaching is also valuable. Magnetic reconnection and geomagnetic induced current are two examples. Magnetic reconnection can be illustrated using a straight current wire and a current loop in magnetism, as demonstrated by the open access MIT materials [20]. The magnetic relaxation in a material such as soft iron could also be illustrated in a LC circuit where the magnetic energy is released/ dumped to the capacitor and resistor. The popular Pasco LCR board has been used. The Gauss-meter recorded pulse shape could be modified by inserting the supplied iron rod into the core region of the coil. Increasing the inductance by more wiring would not alter the pulse shape because the L/R would still be roughly the same. The iron rod insertion would prolong the energy dumping process. The prolonged diffusion profile data of Gauss versus time would be related to the spin alignment decay in the iron rod. The geomagnetic induced current and grounding issues can be part of a discussion in E&M at the introductory physics level with excellent resources [21, 22, 23, 24].

The Money Magazine July 2016 "Value of College" survey showed that students put connectedness as a major reason to get a college education [25, 26]. The perspective of a project acting as a link to the cited references will promote connectedness awareness beyond social media, regardless of the positive or negative results in relationship to the hypothesis. Such a linking perspective is consistent with the prevailing norm in publication practice that a community college can still web-access a published abstract with the cited reference information when the content access is denied due to subscription fee issue. A student writing would link the cited facts in the introduction section and information in the reference section when discussing the results of the analysis. A successful linking would reflect the student's judgment in data interpretation. This experience is no other than the development of critical thinking when Wikipedia explains critical thinking as the linking of facts and information with judgment. The Money Magazine survey also put critical thinking as a high valued issue by students and their parents. The possibility of a community college REU student going onto a senior college REU program will fulfill the pursuit of the number one benefit of college education reported in the Money Magazine survey, that is, the development of a career. Therefore community college REU project provides connectedness awareness, critical thinking and career development, the top three benefits of college education.

Conclusions

REU implementation has been described from the faculty viewpoint in terms of the extraction of information and facts from data mining activity, with feedback to classroom teaching and attention to the "Value of College". Future improvement of our community college REU program could include the offering of AI related research projects [27, 28].

Acknowledgments

We thank Alexei Kisselev for laboratory support. We thank Dr. Chantale Damas for participation in QCC NSF-REU program (PI Lieberman). We thank Eric Cheung MD UIC Medical College Psychiatry Department for discussion. We thank anonymous reviewers for their suggestions.

References

1 LIGO, et al, "GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence" arxiv: 1606.04855 (2016).

2 Fabo Feng, Mikko Tuomi, et al. Color difference makes a difference: four planet candidates around tau Ceti. August 2017. Accepted for publication in AJ. http://lanl.arxiv.org/abs/1708.02051

3 Wilson Tsz-Hon Kowk, et al. Comparative fractal analysis of 2013 November 5 multiple solar eruptions with Fokker-Planck equation using Solar Dynamics Observatory digital images. ASEE 2014 Zone I Conference Proceedings. http://asee-ne.org/proceedings/2014/Student%20Papers/103.pdf

4 NASA IDL Package. https://www.nas.nasa.gov/hecc/support/kb/interactive-data-language-(idl)_119.html

5 Yale University IDL Tutorial. http://exoplanets.astro.yale.edu/tutorials/idl.php

6 NASA Fv software. (Last accessed Sep 1 2017) https://heasarc.gsfc.nasa.gov/ftools/fv/

7 Longcope, D. W., Living Reviews in Solar Physics. Topological Methods for the Analysis of Solar Magnetic Fields, December 2005, 2:7. https://link.springer.com/article/10.12942/lrsp-2005-7

8 Georgios Chintzoglou, Angelos Vourlidas, et al. Magnetic Flux Rope Shredding by a Hyperbolic Flux Tube: The Detrimental Effects of Magnetic Topology on Solar Eruptions, March 2017. Accepted for publication in the Astrophysical Journal, http://lanl.arxiv.org/abs/1706.00057

9 A. F. Rappazzo, W. H. Matthaeus, D. Ruffolo, M. Velli, S. Servidio. Coronal Heating Topology: the Interplay of Current Sheets and Magnetic Field Lines, ApJ 844, 87 (2017). http://lanl.arxiv.org/abs/1706.08983

10 S. A. Mao, C. Carilli, et al. Detection of microgauss coherent magnetic fields in a galaxy five billion years ago. Aug 2017. Accepted Nature Astronomy. http://lanl.arxiv.org/abs/1708.07844

11 Phys.org. Giant magnetic fields in the universe. March 22, 2017 https://phys.org/news/2017-03-giant-magnetic-fields-universe.html

12 Phys.org. Magnetic field discovery gives clues to galaxy-formation processes. June 18, 2015 https://phys.org/news/2015-06-magnetic-field-discovery-clues-galaxy-formation.html

13 Fei Zhao, Heather L. Franco, et al. Elimination of the male reproductive tract in the female embryo is promoted by COUP-TFII in mice. Science Vol 357, Issue 6352, pp 717-720, August 2017. http://science.sciencemag.org/content/357/6352/717

14 Ahsan M. et al. The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. PLoS Genet. 2017 Sep 15;13(9). https://www.ncbi.nlm.nih.gov/pubmed/28915241

15 Stephens NM, Fryberg SA, Markus HR, Johnson CS, Covarrubias R. Unseen disadvantage: how American universities' focus on independence undermines the academic performance of first-generation college students. J Pers Soc Psychol. 2012 Jun;102(6):1178-97. https://www.ncbi.nlm.nih.gov/pubmed/22390227

16 Richard V. Reeves and Joanna Venator. The Inheritance of Education. October 27, 2014. https://www.brookings.edu/blog/social-mobility-memos/2014/10/27/the-inheritance-of-education/

17 Vox.com. The subtle ways colleges discriminate against poor students, explained with a cartoon. Sep 11 2017. https://www.vox.com/platform/amp/2017/9/11/16270316/college-mobility-culture

18 Dunn MC, Bourne PE. Building the biomedical data science workforce. PLoS Biol. 2017 Jul 17;15(7). https://www.ncbi.nlm.nih.gov/pubmed/28715407

19 A. Colin Cameron. EXCEL 2007: Multiple Regression. http://cameron.econ.ucdavis.edu/excel/ex61multipleregression.html

20 G. Li, B. Dasgupta, G. Webb and A. K. Ram. Particle Motion and Energization in a Chaotic Magnetic Field. AIP Conf. Ser. 1183, 201–211 (2009) http://aip.scitation.org/doi/abs/10.1063/1.3266777 (uploaded onto Research Gate by Dasgupta on 23 October 2016. https://www.researchgate.net/publication/233884586)
21 T. Molinski. Why utilities respect geomagnetically induced currents. Journal of Atmospheric and Solar-Terrestrial Physics 64, pp 1765– 1778, 2002. (sciencedirect.com/science/article/pii/S1364682602001268)
22 Denny M. Oliveira, Chigomezyo M. Ngwira. Geomagnetically Induced Currents: Principles. Braz J Phys 47:552–560, 2017. https://link.springer.com/article/10.1007/s13538-017-0523-y
23 Shinichi Watari. Estimation of geomagnetically induced currents based on the measurement data of a transformer in a Japanese power network and geoelectric field observations. Earth, Planets and Space. Vol 67:77 (2015). https://earth-planets-space.springeropen.com/articles/10.1186/s40623-015-0253-8
24 Love, J. J. & Swidinsky, A. Time causal operational estimation of electric fields induced in the Earth's lithosphere during magnetic storms, Geophys. Res. Lett., 41, 2266-2274, (2014). https://geomag.usgs.gov/downloads/publications/10.1002_2014GL059568.pdf.
25 Barnes and Noble College Insight. Today's College Students Value Connections and Experiences More Than Higher Salaries. July 13 2016. http://next.bncollege.com/todays-college-students-value-connections-experiences-higher-salaries/
26 Money Magazine. What makes a college a good value? Parent and Students share similar goals data table: top three benefits of college. August 2016 page 52.
27 Qiantong Xu, Ke Yan, Yonghong Tian. Learning a repression Network for Precise Vehicle Search. August 2017. http://lanl.arxiv.org/abs/1708.02386
28 Yashar D. Hezaveh, Laurence Perreault Levasseur & Philip J. Marshall. Fast automated analysis of strong gravitational lenses with convolutional neural networks. Nature 548, page 555–557. (31 August 2017). http://www.nature.com/nature/journal/v548/n7669/full/nature23463.html

Sunil Dehipawala, PhD

Dehipawala serves as Associate Professor at CUNY Queensborough Community College. His research interests include X-ray absorption, random sequence analysis, and education research.

Raul Armendariz, PhD

Armendariz serves as Assistant Professor at CUNY Queensborough Community College. His interest include high energy physics research, cosmic ray research and education research

George Tremberger, BS

Tremberger serves as Lecturer at CUNY Queensborough Community College.

David Lieberman, PhD

Lieberman serves as Professor and Physics Chair at CUNY Queensborough Community College.

Tak Cheung, PhD

Cheung serves as Professor at CUNY Queensborough Community College.