ASEE 2014 Zone I Conference, April 3-5, 2014, University of Bridgeport, Bridgpeort, CT, USA.

Analysis of Elongation Factor-Tu (EF-Tu) DNA Sequences Using Free Energy & Shannon Entropy

Alessandro DiMarco, France Marquez, Wilson Tsz-Hon Kowk, ShuaiXiang Zhang, Students of Pre-Engineering Program, CUNY Queensborough Community College, Bayside NY 11364

[1]Sunil Dehipawala, [2]Andrew Nguyen, and [1]Tak Cheung, [1]Physics Department , [2]Biology Deapartment, CUNY Queensborough Community College Bayside NY 11364

***Abstract*—A paleo-experimental evolution report on elongation factor EF-Tu structural stability results using the ancestral protein sequence reconstruction modeling has been used in teaching community college pre-engineering students to do research. A project analyzing free energy and Shannon entropy of the engineered DNA sequences would benefit students interested in protein engineering. The studying of EF-Tu DNA sequences obtained from such an ancestral protein sequence reconstruction model shows a clustering pattern in the graph of Shannon entropy versus free energy calculated from the NUPACK software, posted on the internet by Caltech. The inclusion of HB8 and *E. coli* EF-Tu sequences for comparison suggests that a local area on the graph in the unit of bits-kcal/mol would serve as a marker for ancestral protein sequence reconstruction modeling constraints, while the separation distance between two such areas would represent selection pressure differential between organisms. Extension to transcription elongation factor retained the clustering pattern. The inclusion of Pandoravirus supports a regression trend of entropy versus free energy for the studied transcription elongation factor sequences with adjusted R-sq 0.984 (N =4).**

Keywords— Shannon entropy; NUPACK; EF-Tu; transcription elongation factor

I. INTRODUCTION

Community college pre-engineering students need counseling on which career path such as electrical engineering, chemical engineering, protein engineering, etc. Hands-on experience gained in doing a research project in a laboratory and presenting the results in conferences would enhance motivation and improve retention. Genetic circuit engineering, where the proteins from the expression of one gene would regulate the expression of another gene, and protein engineering are relatively new fields where community college pre-engineering students usually have minimal exposure. A very important task in protein engineering would be the study of an amino acid sequence folding with free energy design optimization. Knowing that modern-day proteins were descended from proteins in ancient life forms, the ancestral protein sequence reconstruction technique would offer valuable information. A paleo-experimental evolution report has suggested that the laboratory investigation of the thermo-stability of elongation factor EF-Tu protein by circular dichroism data would provide an opportunity to rewind the tape of life using ancestral protein reconstruction modeling approach [1]. Furthermore they reported that the insertion of an ancient version of EF-Tu in *Escherichia coli* (*E. coli)* would give rise to mutations of other genes while the inserted ancient sequence has remained intact without mutation [2]. The elongation factor thermo unstable (EF-Tu) provides the aminoacyl tRNA access to a free site of the ribosome. Knowing that EF proteins show strong thermo-stability via structural divergence and adapt readily to the host organism operating temperatures, the book of life informatics dogma in biology would point to the correlations of the DNA sequence thermo-adaptability with the associated DNA sequence bioinformatics. The advances in biomedical research have been generating numerous DNA data informatics suitable for analysis with the Shannon entropy formulation in electrical engineering focusing on informatics communication.

The early application of Shannon entropy analysis on DNA sequence, treating a sequence as words, focused on the L-block entropy calculation where L =1 for single alphabet or mono-nucleotide entropy, L = 2 for double-alphabet or dinucleotide pair entropy, etc. [3]. The results for a real DNA sequence, the yeast chromosome sequence, showed maximal block entropy indicative of a very disordered sequence as compared to literary texts or computer codes for the studied L values ranging from 1 to 15. However the empirical distribution of all length-L words shows convergence problems for finite DNA sequences. One of the proposed solutions was to extend the original Shannon formula to Rényi quadratic entropy formula calculated with MATLAB [4]. Recent development of Shannon entropy concept had been summarized in a review volume by Advances in Experimental Medicine and Biology with chapters on motif composition analysis [5], gene expression analysis [6], 3-mer visualization of DNA sequence for pattern analysis [7], and new measures of entropy such as topological entropy [8]. The numerous references cited in the above review volume form a compact literature source for college students using Shannon entropy in bioinformatics related research projects. Among all these entropy concept

applications, our community college student research project focuses on the mono-nucleotide entropy and di-nucleotide entropy correlation with free energy.

In general, a collection of the 16 di-nucleotide pairs of a DNA sequence would yield a histogram representation and an entropy value in informatics. Since a protein could be coded from many different DNA sequences, the imposed constraints would limit the DNA sequence selection in an ancestral protein sequence reconstruction study and useful in studies on "rewind the tape of life" with various investigative hypotheses. This project studied the DNA sequence free energy correlation with entropy informatics in elongation factor EF-Tu.

II. MATERIALS AND METHODS

The Elongation Factor-Tu (EF) DNA sequences used in the paleo-experimental evolution report cited as Reference One were given to us by Professor Eric Gaucher. The *Thermus thermophilus* (*T. thermophilus*) HB8 EF-Tu calibration protein sequence for circular dichroism used in Reference One also has a corresponding DNA sequence in Genbank. Other DNA sequences such as *E. coli* EF-Tu for comparison study in this project are also in Genbank. A genetic DNA sequence can be viewed as an assembly of di-nucleotide pairs. The information content entropy calculations were done using the Shannon (p* log(p)) formulation where p represents the probability of a di-nucleotide pair. Summation over the range of all 16 possibilities or pairs for di-nucleotide entropy would give 4 bits as maximum entropy per possibility. A genetic DNA sequence can also be viewed as an assembly of mono-nucleotides. Note that there are 4 possibilities for mono-nucleotide entropy giving 2 bits per A or T or C or G for maximum entropy. The NUPACK software from Caltech was used to calculate the DNA sequence free energy values.

III. RESULTS OF DNA SEQUENCE ANALYSIS

A typical mono-nucleotide histogram is shown in Figure 1. The correlation of mono-nucleotide entropy with di-nucleotide entropy is shown in Figure 2 for the studied EF-Tu sequences.

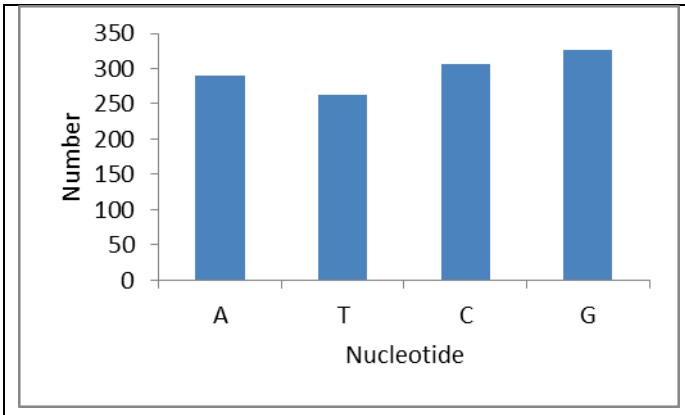

Figure 1: *E. coli* (K-12 substr. MG1655) EF-Tu sequence mono-nucleotide histogram.

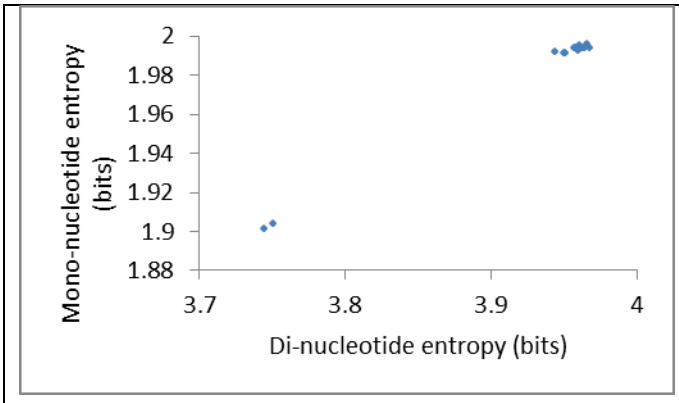

Figure 2: The correlation of mono-nucleotide entropy in bits y-axis with di-nucleotide entropy in bits x-axis is shown for the studied EF-Tu sequences. The *E. coli* EF-Tu sequence clusters with the 12 sequences from ancestral protein reconstruction model of Reference One at the upper right corner while the calibration sequences of HB8 (variant-1 TTHA0251 and variant-2 TTHA1694) used in Reference One cluster at the lower left corner.

The di-nucleotide entropy (x-axis) versus free energy (kcal/mol) computed by NUPACK software from Caltech is displayed in Figure 3 for the studied EF-Tu sequences. The behavior is similar to that of Figure 2 with *E. coli* sequence clusters with the 12 sequences from ancestral protein reconstruction of Reference One at the right while the calibration sequences of HB8 (variant-1 TTHA0251 and variant-2 TTHA1694) used in Reference One cluster at the left.

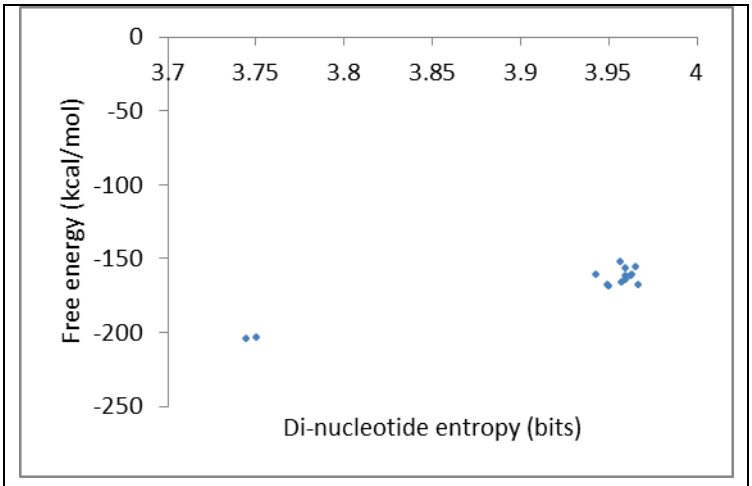

Figure 3: The di-nucleotide entropy in bits (x-axis) versus free energy (kcal/mol) for the studied EF-Tu sequences. The HB8 sequence data (used for calibration in Reference One) cluster at the lower left corner.

IV. DISCUSSION

In general, the 12 reconstructed DNA sequence in the ancestral protein reconstruction of Reference One cluster with *E. coli* DNA sequence while the HB8 DNA sequences responsible for the EF-Tu proteins used for circular dichroism data calibration does not. The observed clustering effect in Figure 2 would suggest that the studied HB8 sequences having lower entropy values could have been subjected to relatively strong selection pressure that moved them away from maximal disordered states, consistent with the entropy interpretation offered in Reference Three. An ancient sequence operating at a high temperature would be expected to have a strong free energy as computed with NUPACK software from Caltech. Reference One reported the use of EF-Tu from modern-day *T. thermophilus* (HB8) in their laboratory calibration (76.7 C circular dichroism data). However the earliest EF-Tu protein sequence in the ancestral protein reconstruction model at 73.3 Celsius in Reference One has a weak free energy value in the associated DNA sequence of -156 kcal/mol (3.961 bits in entropy, Figure 3) which is much lower than the two HB8 counterparts at around (-200 kcal/mol, 3.75 bits). Ancestral protein sequence reconstruction aims to create a phylogenetic tree and conjecture ancestral amino acid states at nodes. The fact that there could be an overestimation in properties such as thermo-stability prediction is acceptable [9], provided that there would be further supporting evidence such as circular dichroism on the structural folding, as demonstrated in Reference One. Another alternative approach would be the use of ancestral gene sequence reconstruction to supplement the ancestral protein sequence reconstruction so as to rewind the tape of life [10]. Recent suggestions include the modeling of the evolution of a gene under selective pressure with marginally stability for protein sequences [11].

It would be instructive to compare translation elongation factor to transcription elongation factor, and to expand the range of entropy and free energy values and to search for a broad understanding. In particular a virus would have transcription elongation factor with no need for translation elongation factor. The recently discovered large genome Pandoravirus carries transcription elongation factor S-II with sequence data catalogued in Genbank as *Pandoravirus salinus* partial genome KC977571 and *Pandoravirus dulcis* complete genome KC977570 [12, 13]. The HB8 and *E. coli* transcription elongation factors were included in the comparison. The result is displayed in Figures 4 and 5.

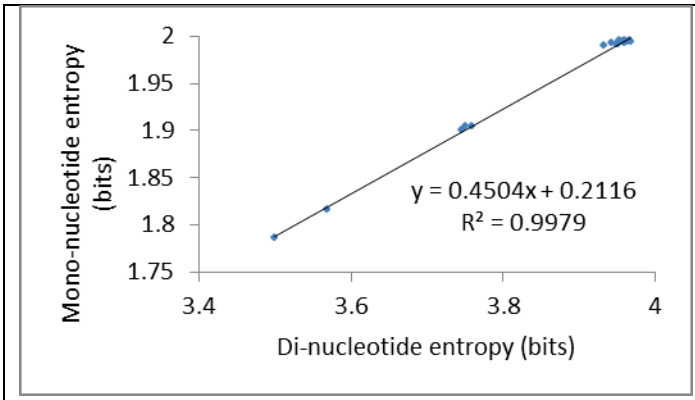

Figure 4: The correlation of mono-nucleotide entropy in bits (y-axis) with di-nucleotide entropy in bits (x-axis) is shown (N = 19).

In Figure 4, the two Pandoravirus transcription elongation sequences cluster at the lower corner, the HB8 transcription elongation factor (nsuA also labeled as TTHA0701) clusters with its translation elongation factor (2 sequences) at the middle, and the *E. coli* transcription elongation factor nsuA clusters with its translation elongation factor and Reference One sequences at the upper corner. The high R-sq value of 0.99 (N = 19) shows that the Shannon entropy measure of information content is similar for mono-nucleotide types or di-nucleotide pairs. A similar clustering pattern for entropy and free energy for the studied sequences is shown in Figure 5.

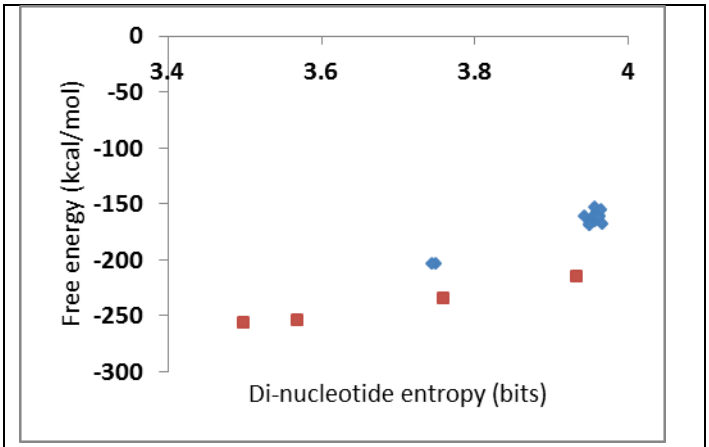

Figure 5: The di-nucleotide entropy in bits (x-axis) versus free energy (kcal/mol) for the studied EF-Tu translation (diamonds) and transcription (squares) elongation factor sequences. The *E. coli* nusA transcription elongation factor sequence (3.933 bits, -215 kcal/mol) clusters with the *E. coli* translation EF-Tu sequence and the 12 sequences from ancestral protein reconstruction model of Reference One at the upper right corner while the calibration sequences of HB8 EF-Tu (variant-1 TTHA0251 and variant-2 TTHA1694) used in Reference One cluster at middle together with TTHA0701 or nusA transcription elongation factor (3.759 bits, -234.4 kcal/mol). The Pandoravirus transcription elongation factor sequences (diamonds) cluster at the lower left corner.

It could be a speculative interpretation that the clustering feature displayed in Figure 4 shows virus transcription elongation factors must have low entropy values with relatively ordered sequences as compared to the translation elongation factor sequences in general. A broad understanding emerges where entropy and free energy are clustered in the studied translation and transcription elongation factor sequences as displayed in Figure 5. The model dependent ancestral protein reconstruction of EF-Tu sequences of Reference One cluster together with the *E. coli* EF-Tu sequence near the upper corner, and the occupied area could serve as a marker for model constraints. The HB8 EF-Tu sequences cluster together with natural selection constraints at the middle, and the Pandoravirus having only transcription elongation factor has the most constraint in terms of entropy informatics and its sequences cluster at the lower corner. The local neighborhood with the unit of bits-kcal/mol would serve as a comparative parameter representing the amount of constraints, and the separation between two such neighborhoods or areas would represent selection pressure differential between organisms. Interestingly, the three different constraints as discussed above show some correlation with free energy values, suggesting that entropy does not distinguish the constraint origin. Specifically, the studied transcription elongation factor shows a high correlation (adjusted R-sq 0.984 N = 4) for entropy with free energy in Figure 6. Transcription carries a complex multi-step process where RNA polymerase II (Pol II) would transcribe DNA informatics into RNA with timing control including transcription initiation, elongation, capping, termination, and histone modifying factors. The high R-sq value would be expected to decrease when other transcription elongation factors from other organisms are included in future studies. In general, a sequence having relatively no constraint in information content would carry high entropy value near the maximum allowable 4 bits for di-nucleotide and 2 bits for mono-nucleotide. A high CG or AT content sequence would carry low entropy value with a narrow histogram as compared to the histogram of a sequence with equal probability for each di-nucleotide or mono-nucleotide. However a high CG content sequence would carry strong free energy (lower left corner of Figure 6) and be favorable to form a positive slope regression with other sequences as shown in Figure 6. On the other hand, a high AT sequence would carry low free energy (upper left corner of Figure 6) and be favorable to form a negative slope regression with sequences near maximum entropy and moderate free energy. The high AT sequence conjecture could be solved with future investigations on sequences from other organisms. Whether selection would favor high AT or CG content in transcription (or translation) elongation factors among organisms could be another interesting question for future studies.

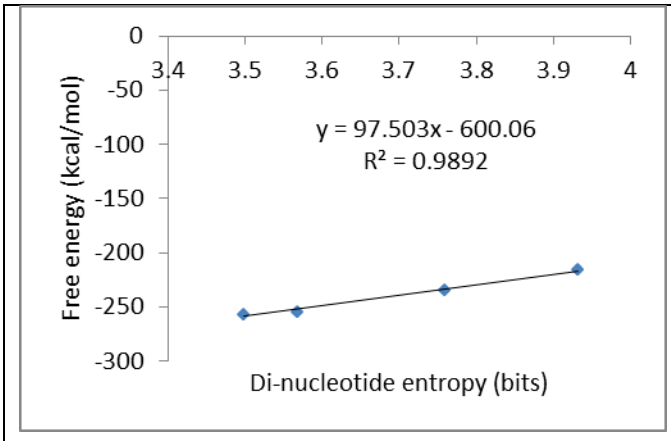

Figure 6: The di-nucleotide entropy in bits (x-axis) versus free energy (kcal/mol) for the studied transcription elongation factor sequences (adjusted R-sq ~ 0.984).

The entropy values of transcription elongation factor are closed to translation elongation factor in the studied HB8 and *E. coli* sequences (Figure 5). This closeness could be a universal signature saying that both transcription and translation elongation factors had experienced similar constraints. The HB8 *T. thermophilus* was originally isolated from a hot spring thermal vent and the lower entropy and stronger free energy in its elongation factor sequences as compared to the *E. coli* counterparts would be consistent with a trend that special environment for strong free energy sequence would impose more constraints and thus would suppress information content entropy. The Pandoravirus, having CpG content of about 20% and with only 6% sequence similarity with all other sequences from all known viruses, would follow the same pattern in its transcription elongation factor of low entropy and strong free energy and occupies the lower corner of the entropy-free energy graph. Knowing that ChIP with DNA microarray coupled with high-throughput sequencing technologies (ChIP-chip and ChIP-seq) have revealed wide spread regulation of transcription elongation [14], and that stochastic fluctuation in gene expression has been interpreted as decision making or free will in bacteria [15], the proposed entropy-free energy approach would offer additional insights.

V. CONCLUSIONS

The studying of EF-Tu DNA sequences obtained from an ancestral protein reconstruction model shows a clustering pattern in the graph of Shannon entropy versus free energy calculated from the NUPACK software. Extension to transcription elongation factor retained the clustering pattern. The inclusion of Pandoravirus supports a regression trend for transcription elongation factor with adjusted R-sq 0.984 (N = 4). Future studies could include the entropy-free energy analysis of all the DNA sequences in the recently discovered Pandoraviruses (amoebic virus), having only 6% sequence similarity with all other sequences from all known viruses, for the examination of the already proposed hypothesis that

Pandoraviruses could have some sequences from non-Earth DNA source.

ACKNOWLEDGMENT

We thank Professor Eric Gaucher for discussion and the sharing of the DNA data from his laboratory. We thank NUPACK staff for the software and its internet availability. We thank the research groups for posting the gene data from their laboratories in the public domain.

REFERENCES

[1] Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 2008 Feb 7;451(7179):704-7.

[2] Georgia Tech News Release , "Giving Ancient Life Another Chance to Evolve", Posted July 11, 2012 Atlanta, GA , Georgia Tech News Release July 11 2012 : " …….After achieving the difficult task of placing the ancient gene in the correct chromosomal order and position in place of the modern gene within *E. coli*, Kacar produced eight identical bacterial strains and allowed "ancient life" to re-evolve. This chimeric bacteria composed of both modern and ancient genes survived, but grew about two times slower than its counterpart composed of only modern genes. "The altered organism wasn't as healthy or fit as its modern-day version, at least initially," said Gaucher, "and this created a perfect scenario that would allow the altered organism to adapt and become more fit as it accumulated mutations with each passing day." The growth rate eventually increased and, after the first 500 generations, the scientists sequenced the genomes of all eight lineages to determine how the bacteria adapted. Not only did the fitness levels increase to nearly modern-day levels, but also some of the altered lineages actually became healthier than their modern counterpart. When the researchers looked closer, they noticed that every EF-Tu gene did not accumulate mutations. Instead, the modern proteins that interact with the ancient EF-Tu inside of the bacteria had mutated and these mutations were responsible for the rapid adaptation that increased the bacteria's fitness. In short, the ancient gene has not yet mutated to become more similar to its modern form, but rather, the bacteria found a new evolutionary trajectory to adapt……" http://www.gatech.edu/newsroom/release.html?nid=138621

[3] Schmitt AO, Herzel H. "Estimating the entropy of DNA sequences." J Theor Biol. 1997 Oct 7; 188(3), pp369-77.

[4] Vinga S1, Almeida JS. "Rényi continuous entropy of DNA sequences." J Theor Biol. 2004 Dec 7; 231(3), pp377-88.

[5] Stojanovic N, Singh A. "Exploring motif composition of eukaryotic promoter regions." Adv Exp Med Biol. 2010;680, Chapter 4 (pp27-34).

[6] Liu GG, Fong E, Zeng X. "GNCPro: navigate human genes and relationships through net-walking." Adv Exp Med Biol. 2010; 680, Chapter 29 (pp253-9).

[7] Cox DN, Tharp AL." Toward a visualization of DNA sequences."Adv Exp Med Biol. 2010;680, Chapter 48 (pp419-35).

[8] Garrido A. "Some new measures of entropy, useful tools in bio-computing." Adv Exp Med Biol. 2010;680, Chapter 83 (pp745-50).

[9] Williams PD, Pollock DD, Blackburne BP, Goldstein RA. "Assessing the accuracy of ancestral protein reconstruction methods." PLoS Comput Biol. 2006 Jun 23;2(6):e69.

[10] Harms MJ and Thornton JW. "Analyzing protein structure and function using ancestral gene reconstruction." Harms, "Curr Opin Struct Biol. 2010 Jun;20(3):360-6

[11] Goldstein RA., "The evolution and evolutionary consequences of marginal thermostability in proteins", Proteins. 2011 May;79(5):1396-407.

[12] Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie JM, Abergel C. , "Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes", Science. 2013 Jul 19;341(6143):281-6.

[13] Wind M and Reines D., "Transcription elongation factor SII.", Bioessays. 2000 Apr;22(4):327-36.

[14] Gilchrist DA, Fargo DC, Adelman K., "Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation.", Methods. 2009 Aug;48(4):398-408.

[15] Kondev, Jane, "Bacterial decision making.", Physics Today 67(2), 31 (2014)

Biography

Mr. Alessandro DiMarco is a pre-engineering student and he is interested in chemical engineering.

Ms. France Marquez is a pre-engineering student and she is interested in biomedical engineering.

Mr. Wilson Tsz-Hon Kowk is a pre-engineering student and he is interested in information technology engineering.

Mr. ShuaiXiang Zhang is a pre-engineering student and he is interested in software engineering.

Dr. Sunil Dehipawala is a professor of physics and his experiences include Synchrotron based spectroscopy, pedagogy, etc. (sdehipawala@qcc.cuny.edu)

Dr. Andrew Nguyen is a professor of biology and his experiences include genetics, pedagogy, etc. (anguyen@qcc.cuny.edu)

Dr. Tak Cheung is a professor of physics and his experiences include thin film physics, pedagogy, etc. (tcheung@qcc.cuny.edu)