A Structural Equation Model Study of Shannon Entropy Effect on CG content of Thermophilic 16S rRNA and Bacterial Radiation Repair Rec-A Gene Sequences

T. Holden, P. Schneider, E. Cheung, J. Prayor, R. Duran, J. Ye, G. Tremberger Jr, D. Lieberman & T. Cheung

CUNY Queensborough Community College Physics and Biology Departments
222-05 56th Ave Bayside NY 11364

Abstract

This project studied the latent variables in datasets of 16S rRNA gene sequences from 9 thermophiles and Rec-A radiation repair genes from 5 bacteria in terms of the Shannon entropy, fractal dimension, CG content, and optimum growth temperature or radiation dosage. Gene sequence nucleotide fluctuation could be studied in terms of Shannon entropy and fractal dimension, when the nucleotide atomic numbers are treated as data in a random series. Previous studies reported that Shannon entropy, sequence CG content, fractal dimension, and optimum growth temperature for the 16S rRNA or radiation dosage for the Rec-A sequences are correlated. Shannon entropy correlation was found to have an R-square of 0.80 (N = 9) with optimum growth temperature in the studied 16S rRNA sequences, and R-square of 0.94 (N = 5) with radiation dosage in the studied Rec-A sequences, respectively. This project found that Shannon entropy and fractal dimension are indicators of an arbitrary latent variable, which could contain a bioinformatics signal ("code-bioinformatics"), and that the CG content and optimum growth temperature (or radiation dosage) are indicators of another latent variable, which could be related to adaptation ("adapt-aptitude"). The structural equation model suggests that CG content and optimum growth temperature or radiation dosage could be adaptation based while the Shannon entropy and fractal dimension could be code-bioinformatics based. This information could support genetic engineering strategies and the presented SEM methodology could be applied to sequences with environmental parameters other than temperature and radiation dosage. This student project's SEM calculations were done with the free student version of LISREL, which is available to students majoring in the biological and health sciences as well. A simple lizard projectile motion numerical example is provided for educational purpose such that the interested readers can verify their operation of the free student version of LISREL.

Keywords: Shannon entropy, structural equation model, LISREL software, bioinformatics, fractal dimension, lizard motion data for educational purpose

Introduction

Structural equation models (SEM) have been used to probe the aspects that concern causative hypotheses/elements contained in datasets [1, 2]. The causative hypotheses/elements would convey causal assumptions, but not necessarily a model that

would generate validated causal conclusions. The acceptance and consistency within a SEM statistical calculation does not really prove the causal relationships in a model Structural equation model analysis is an improvement over correlation relationship analysis and points toward potential causal relationship. LISREL (LInear Structural RELations) is a popular software package used by researchers for structural equation modeling [3]. The methodology has been most popular in the subfield of gene expression in genetic research. Gene sequence nucleotide fluctuation could be studied in terms of Shannon entropy; and fractal dimension when the nucleotide atomic numbers are treated as data in a random series [4]. Shannon entropy has been accepted as a measurement of the information content while position sensitive fractal dimension could be a result of adaptation or entropy related. This explorative study focuses on the classification of Shannon entropy, fractal dimension, CG (C + G) content and optimum growth temperature or radiation dosage in terms of bioinformatics and adaptation.

Previous studies reported that Shannon entropy, sequence CG content, fractal dimension, and optimum growth temperature for the 16S rRNA or radiation dosage for the Rec-A sequences are correlated [5, 6].

The Shannon entropy and sequence CG content were found to correlate with optimal growth temperature, having an R-square of 0.8 (negatively), and 0.88 (positively), respectively, for 9 reported thermophilic 16S rRNA sequences. The fractal dimension was also found to be correlated (positive) to the sequence CG content with an R-square value of 0.74.

The Shannon entropy and sequence CG content were found to correlate with radiation dosage, having an R-square of 0.94 (negatively), and 0.81 (positively), respectively, for 5 reported bacterial Rec-A sequences. The fractal dimension was also found to be correlated (positively) to the sequence CG content with an R-square value of about 0.63.

Furthermore the Shannon entropy was found to be correlated with fractal dimension for the studied 16S rRNA sequences (positively with R-square ~ 0.82), and Rec-A sequences (negatively with R-square ~ 0.66 N = 5) respectively.

Data

The 16S rRNA and Rec-A gene sequences were downloaded from Genbank with accession numbers provided in the literature [5, 6].

The studied 16S rRNA sequences were from thermophiles. The studeied archaea-euryarchaeota thermophiles were *archaeoglobus fulgidus*, *methanothermobacter thermautotrophicus*, *methanocaldococcus jannaschii*, *pyrococcus horikoshii* and *thermoplasma acidophilum*. The studied aracheae-crenarchaeota thermophiles were *Sulfolobus solfataricus* and *aeropyrum pernix*. The studied bacterial thermophiles were *thermotoga maritime* and *aquifex aeolicus*.

The studied Rec-A radiation repair sequences were from bacteria. The studied bacteria were *Deinococcus radiodurans*, *Deinococcus geothermalis*, *E coli K-12*, *Pseudomonas putida KT2440* , and *Shewanella oneidensis MR-1.*

The Shannon entropy was calculated using the probability of the 16 di-nucleotide pairs. The fractal dimension was calculated using the Higuchi method [4]. The structural equation models were fitted via the free student version of LISREL.

Models

The path diagram of the structural equation model for the studied 16S rRNA and Rec-A sequences are displayed in Figure 1 and 2 with the numeric results respectively.
The Shannon entropy and fractal dimension are indicators for an arbitrary latent variable, which could contain a bioinformatics signal ("code-bioinformatics"), and that the CG content and optimum growth temperature in figure 1 (or radiation dosage in logarithm of Gyr in figure 2) are indicators for another latent variable, which could be related to adaptation ("adapt-aptitude"). The structure equation model suggests that CG content and optimum growth temperature in Figure 1 or radiation dosage ln-Gy in Figure 2 could be adaptation based while the Shannon entropy and fractal dimension could be code-bioinformatics based. The negative coefficient in Figure 2 would be consistent with the previous report that Shannon entropy was found to be negatively correlated with fractal dimension in the studied Rec-A sequences [6].

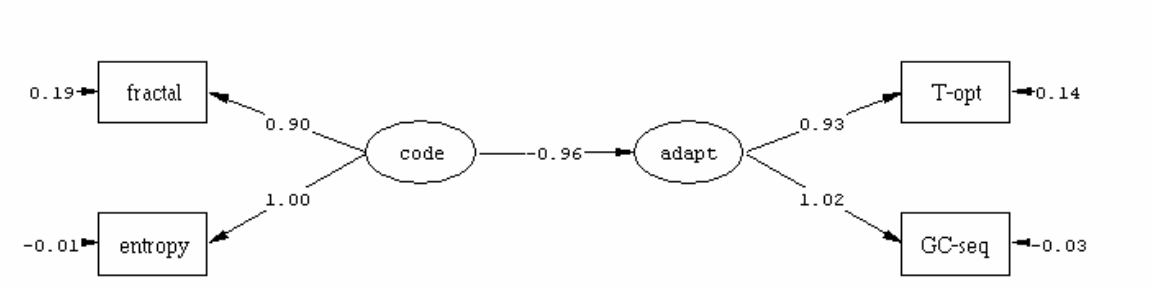

Figure 1: Path diagram of the structural equation model for the studied 16S rRNA sequences

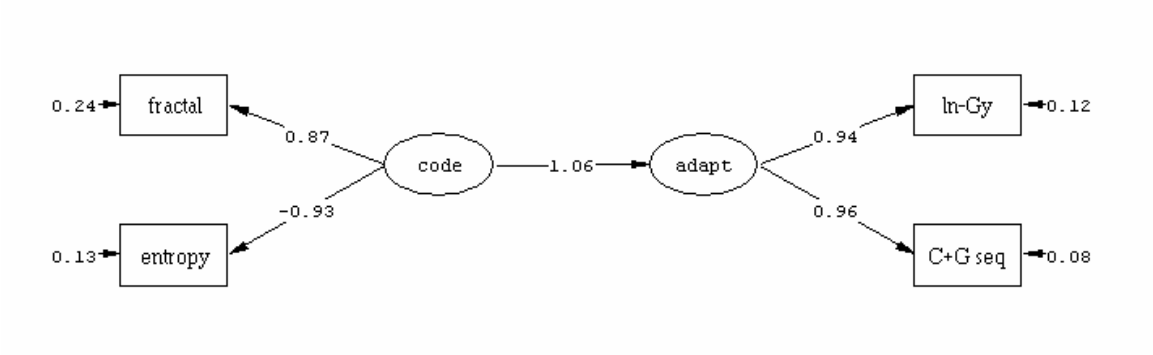

Figure 2: Path diagram of the structural equation model for the studied Rec-A sequences

Discussion

Failed non-converging models are not presented in this report. For example, the grouping of CG content as indicator for the latent variable "code-bioinformatics" would fail to converge in the SEM calculations. The so called failures actually could reveal misconceptions and offer insights on the code-bioinformatics and adapt-aptitude latent variables. Both models would give consistent numeric results when the optimum growth temperature or radiation dose is replaced by the sequence base pair numbers. It would be interesting to apply the analysis to an arbitrary dataset. The Y-chromosome has 429 identified genes so far. A subset of gene sequences with good correlations among the fractal dimension and entropy, sequence base pairs and CG content have failed to give a converging solution as far as we know. It would appear that the external environment inputs such as optimum growth temperature, radiation dosage, etc be important parameters.

A data example is provided below for the interested readers to verify their operation of the free student version of LISREL. The Table 1 data are based on the lizard projectile motion video published by Berkeley Tailbot Robot Engineering in open access format [7]. The numerical values are approximations obtained from the digitization of the video and are provided here for educational purpose.

Table 1: Lizard motion data: frame refers to the frame number, which is the timing and/or horizontal position (arbitrary unit). The tail-angle, body-angle, and height (relative vertical position) indicators are in arbitrary units.

frame	tail-angle	body-angle	height
1	169.5	163.6	65
2	161.1	151.9	70
3	168.8	153.5	64
4	162.3	158.7	67
5	142.7	163.1	75
6	145	161	78
7	171.7	167.2	80
8	147.1	174.2	92
9	145.1	177	95
10	147.7	174.7	98
11	151.6	170.7	103
12	155.7	171.6	104
13	153.7	173	104
14	154.1	172.5	108
15	148.5	156.3	108

A path analysis of Table 1 data was performed with the LISREL numerical results displayed in Figure 3. The interpretation of the latent variable "abc" could include "angular momentum control aptitude" etc, and the latent variable "xyz" could include "adapt-to-gravity-aptitude" etc.

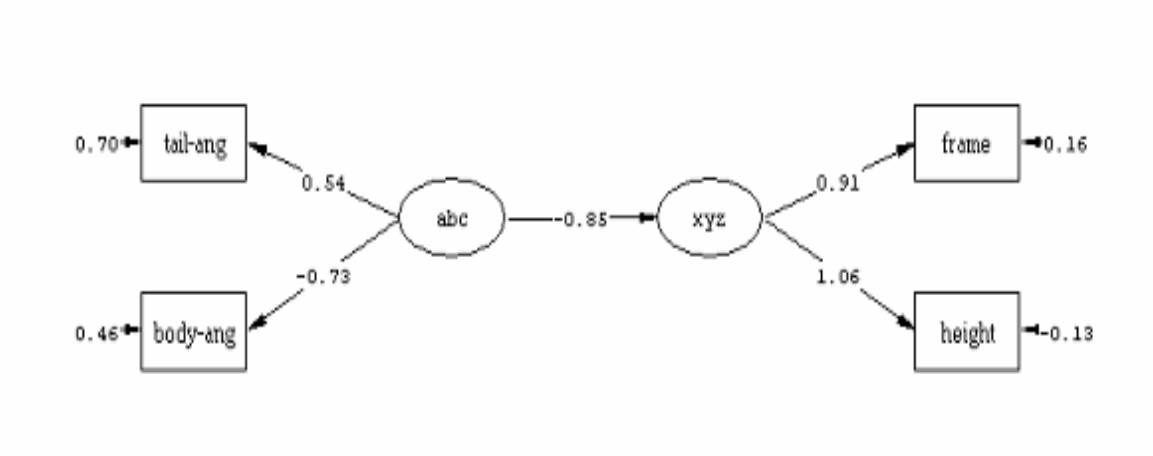

Figure 3: Path Diagram of the structural equation model for the lizard motion data. The LISREL basic model and standardized solution options were selected.

This student project's SEM calculations were done with the free student version of LISREL, which is available to students majoring in the biological and health sciences as well. The SEM analysis would support genetic engineering strategies that focus on code-bioinformatics approach [8].

Conclusions

This explorative project shows that structural equation model is capable of providing quantitative causative hypotheses/elements information on latent variables which could be related to bioinformatics and adaptation respectively. The presented SEM methodology could be applied to sequences with environmental parameters other than temperature and radiation. The lizard projectile motion educational example will help interested readers to verify their operation of the free student version of LISREL.

Acknowledgements

NIH-RIMS Grant (PI: Schneider) and NSF-REU Grant (PI: Lieberman) are gratefully acknowledged. EC, JP, RD and JY thank Professors Holden, Schneider, Tremberger, Jr, Lieberman and Cheung for guidance.

References

1. Beran TN, Violato C. (2010)," Structural equation modeling in medical research: a primer", BMC Res Notes. 2010 Oct 22;3:267.

2. Bollen KA, Noble MD 2011, "Structural equation models and the quantification of behavior", Proc Natl Acad Sci U S A. 2011 Sep 13;108

3. LISREL software
http://www.ssicentral.com/

4. T. Higuchi, "Approach to an irregular time series on the basis of fractal theory", Physica D, vol 31, 277-283, 1998.

5. Todd Holden, G. Tremberger, Jr, E. Cheung, R. Subramaniam, R. Sullivan, N. Gadura, P. Schneider, P. Marchese, A. Flamholz, T. Cheung, and D. Lieberman (2008) Fractal analysis of 16S rRNA gene sequences in archaea thermophiles. Fifth International Conference on Bioinformatics, Computational and Systems Biology, Proc World Acad Sci Engr &Tech, Vol.44, Vol. 44, p31-35, ISSN: 2070-3724

6. Todd Holden, R. Subramaniam, R. Sullivan, E. Cheung, C. Schneider, G. Tremberger, Jr., A. Flamholz, D. H. Lieberman, and T. D. Cheung, (2007) "ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes", Proc. SPIE Vol 6694, 669417

7. http://spectrum.ieee.org/automaton/robotics/diy/uc-berkeley-tailbot-robot-with-tail-shows-off-more-midair-skills (last accessed April 6 2012)

8. Stein CM, Morris NJ, Nock NL (2012), "Structural equation modeling", Methods Mol Biol. 2012;850:495-512.