

# Fractal Analysis of 16S rRNA Gene Sequences in Archaea Thermophiles

T. Holden, G. Tremberger, Jr, E. Cheung, R. Subramaniam, R. Sullivan, N. Gadura, P. Schneider, P. Marchese, A. Flamholz, T. Cheung, and D. Lieberman

**Abstract**—A nucleotide sequence can be expressed as a numerical sequence when each nucleotide is assigned its proton number. A resulting gene numerical sequence can be investigated for its fractal dimension in terms of evolution and chemical properties for comparative studies. We have investigated such nucleotide fluctuation in the 16S rRNA gene of archaea thermophiles. The studied archaea thermophiles were *archaeoglobus fulgidus*, *methanothermobacter thermautotrophicus*, *methanocaldococcus jannaschii*, *pyrococcus horikoshii*, and *thermoplasma acidophilum*. The studied five archaea-eyryarchaeota thermophiles have fractal dimension values ranging from 1.93 to 1.97. Computer simulation shows that random sequences would have an average of about 2 with a standard deviation about 0.015. The fractal dimension was found to correlate (negative correlation) with the thermophile's optimal growth temperature with  $R^2$  value of 0.90 ( $N=5$ ). The inclusion of two aracheae-crenarchaeota thermophiles reduces the  $R^2$  value to 0.66 ( $N=7$ ). Further inclusion of two bacterial thermophiles reduces the  $R^2$  value to 0.50 ( $N=9$ ). The fractal dimension is correlated (positive) to the sequence GC content with an  $R^2$  value of 0.89 for the five archaea-eyryarchaeota thermophiles (and 0.74 for the entire set of  $N=9$ ), although computer simulation shows little correlation. The highest correlation (positive) was found to be between the fractal dimension and di-nucleotide Shannon entropy. However Shannon entropy and sequence GC content were observed to correlate with optimal growth temperature having an  $R^2$  of 0.8 (negative), and 0.88 (positive), respectively, for the entire set of 9 thermophiles; thus the correlation lacks species specificity. Together with another correlation study of bacterial radiation dosage with RecA repair gene sequence fractal dimension, it is postulated that fractal dimension analysis is a sensitive tool for studying the relationship between genotype and phenotype among closely related sequences.

**Keywords**—Fractal dimension; archaea thermophiles; Shannon entropy; GC content

## I. INTRODUCTION

A standard tool in evolutionary biology is genome comparison. The ATCG nucleotide changes over a gene sequence can be viewed as a fluctuation and consequently, can be analyzed with standard tools that include correlation and fractal dimension. For this study, the numerical sequence

representing the fluctuation of ATCG nucleotides in a gene sequence was generated using the atomic number of each element in a nucleotide [1]. This numerical series can then be further processed using methods such as a moving average, which is often used in stock market time series analysis. The resulting fractal dimension of this random series or random series derived from the original atomic number based sequence can be computed. Nucleotide fluctuation has been studied using other assignment schemes [2, 3, 4]. The use of proton number was motivated partly by the observation of mass fractal dimension in the X-ray data of proteins and ribosomes [5], and using a proton assignment scheme may reveal proton sensitivity in the underlying genetic sequence to the folding induced mass fractal. The project aims to use fractal dimension for comparing the bioinformatics on the sequences so that the choice of an assignment scheme would only have a secondary effect. A recent comparison of human and chimpanzee genomes revealed that it is possible to measure the acceleration rate of the accelerated regions of the human genome [6]. The most accelerated region, HAR1, was shown by a gene expression experiment in the human embryo to be transcription active and co-expressed with reelin, which is an essential protein involved in the development of the six-layer cortex of the human brain. Fractal analysis was applied to the HAR1 nucleotide sequence and the homologous sequence in the chimpanzee genome. Analysis shows that the differences in fractal dimension can be used as a marker of evolution [1]. The 118-bp region in HAR1 contains 18 points of substitutions over an evolutionary span of 5 million years when comparing the human to the chimpanzee. However, the same 118-bp region only contains two points of substitutions over a span of 300 million years when comparing the chicken to the chimpanzee. The implications of evolution and positive selection have been discussed in recent literature [7]. The project had used various assignment schemes on the HAR1 RNA sequence [1]. The human sequence was found to have consistently higher fractal dimension than the chimp sequence regardless of the numerical assignment scheme employed. Fractal dimension can be a measure of the capacity dimension, which is the upper bound for information dimension [8]. Which particular assignment scheme would show the largest fractal dimension difference between two sequences, although important, but was not a focus of the current study.

This project focused on the archaea thermophiles with

T. Holden, G. Tremberger, Jr, E. Cheung, P. Marchese, A. Flamholz, T. Cheung, and D. Lieberman are with CUNY Queensborough Community College, Physics Department, Bayside, NY 11364 USA (corresponding author: Todd Holden, email: tholden@qcc.cuny.edu).

R. Subramaniam, R. Sullivan, N. Gadura, and P. Schneider are with CUNY Queensborough Community College, Biology Department, Bayside, NY 11364 USA.

optimal growth temperatures between 60 and 90 degree Celsius. The nucleotide content in the stem structure and hairpin loop structure of the 16S rRNA gene was found to be proportional to the optimal growth temperature [9]. Despite its success, nucleotide content statistics (and also Shannon entropy content) are position independent and lack species specificity information. In contrary fractal dimension, being position sensitive is a useful probe for species specificity. The ability of thermophiles to live at high temperatures would shed light on the conjecture that some extremophiles might have been able to survive in other planets. The project had investigated the nucleotide fluctuation in the 16S rRNA genes of archaea thermophiles as fractal forms. Using the human HAR1 and chimpanzee comparison result, inferences on the evolution and positive selection of the studied archaea thermophiles can be drawn.

## II. MATERIALS & METHODS

### A. Genetic Sequence

The 16S rRNA gene sequences were downloaded from Genbank. The studied archaea-euryarchaeota thermophiles were *archaeoglobus fulgidus* with Accession NC\_000917 Region: complement 1788987..1790478, *methanothermobacter thermautotrophicus* with Accession NC\_000916 Region: 1718787..1720265, *methanocaldococcus jannaschii* with Accession NC\_000909 Region: 638452..639929, *pyrococcus horikoshii* with Accession NC\_000961 Region: 190975..192469, and *thermoplasma acidophilum* with Accession NC\_002578 Region: complement 1474300..1475770. The studied archaea-crenarchaeota thermophiles were *Sulfolobus solfataricus* with Accession NC\_002754 Region: 871672..873167, and *aeropyrum pernix* with Accession NC\_000854 Region: complement 1218714..1220913. Note that *aeropyrum pernix* has an embedded intron so the exon at join (1..878, 1578..2200) was used in this study. The studied bacterial thermophiles were *thermotoga maritime* with Accession NC\_000853 Region: 188968..190526 and *aquifex aeolicus* with Accession NC\_000918 Region: 1192069..1193655.

### B. Higuchi Fractal Method

Among the various fractal dimension methods, the Higuchi fractal method is well suited for studying signal fluctuation [10] and has been applied to nucleotide sequences [11]. In this study, the ATCG sequence was converted to a numerical sequence by assigning the atomic number, the total number of protons, in each nucleotide: A(70), T(66), C(58), G(78). The assigned number is proportional to the nucleotide mass (ignoring isotopes). The A-T and C-G pairs in double stranded DNA have the same value of 136. The numerical sequence I could be used to generate a difference series  $(I(j)-I(i))$  for different lags. The non-normalized apparent length of the series curve is simply  $L(k) = \sum \text{absolute}(I(j)-I(i))$  for all  $(j-i)$  pairs that equal to  $k$ . The number of terms in a  $k$ -series varies and normalization must be used. The normalization is in open literature [12]. If the  $I(i)$  is a fractal function, then the log

$(L(k))$  versus  $\log(1/k)$  should be a straight line with the slope equal to the fractal dimension. Sometime  $\ln(L(k))$  vs  $\ln(1/k)$  can be used as well [13]. Higuchi incorporated a calibration division step (divided by  $k$ ) such that the maximum theoretical value is calibrated to the topological value of 2. When comparing the dimension of two fractal forms, the popular method of taking the difference of the two Higuchi fractal dimension values is valid to within a constant regardless of the calibration division step. The Higuchi fractal algorithm used in this project was calibrated with the Weierstrass function. This function has the form  $W(x) = \sum a^{-nh} \cos(2\pi a^n x)$  for all the  $n$  values 0, 1, 2, 3... The fractal dimension of the Weierstrass function was given by  $(2-h)$  where  $h$  takes on an arbitrary value between zero and one.

The HAR1 and its chimpanzee counterpart sequences are used to illustrate the Higuchi method. The fractal dimension of the 118-bp region HAR1 sequence with atomic number as the numerical values was shown to be about 2.02 for human (Fig. 1), and about 1.97 for chimpanzee, a difference of about 0.05. In Fig. 1, the first seven points were used to calculate the slope. The complement sequence (A becomes T, C becomes G and vice versa) gave the same fractal dimensions for human and chimpanzee, respectively.

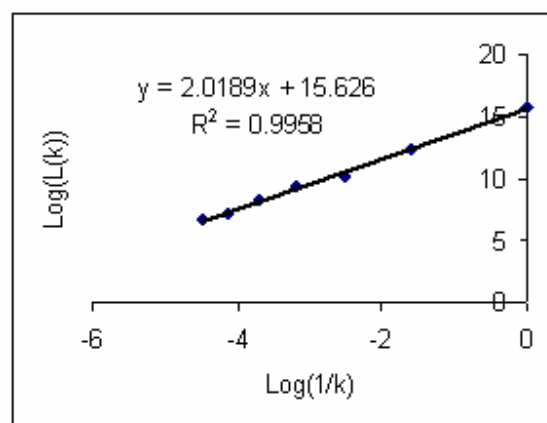


Fig. 1 The fractal dimensions of the human HAR1 sequences for the substitution of 70, 66, 58, 78 for A, T, C, G, respectively. The sequence has 118 nucleotides

The fractal dimension difference was investigated using computer generated random sequences. A 118-data random series with data entry values of 70, 66, 58, and 78 was generated and the fractal dimension distribution was calculated. The GC content was set at the human 42% level. A fractal dimension distribution for 1000 simulated sequences is shown in Fig. 2.

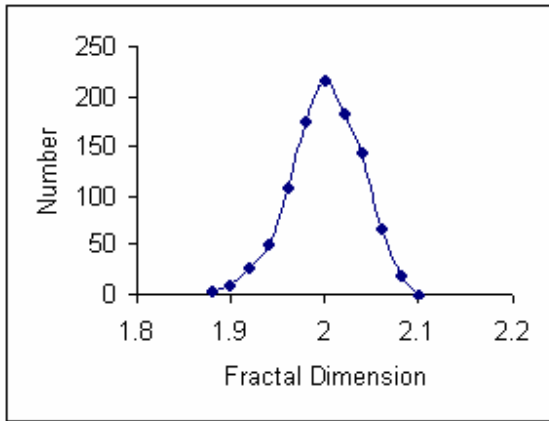


Fig. 2 Distribution of computer-generated random sequence fractal dimension (N = 1000). The average was 2.00 and the standard deviation was 0.04

The distribution has an average of about 2.00 and a standard deviation of 0.04. The simulations showed that the fractal dimension difference between human and chimpanzee was about one standard deviation, given a short sequence of 118 data points.

### III. RESULTS AND DISCUSSION

#### A. Fractal Analysis

The fractal dimension versus T-opt (the optimal growth temperature) relationship for the studied five archaea-eyryarchaeota thermophiles is displayed in Fig. 3. The T-opt information for the 9 studied organisms in this project was obtained from Reference 9.

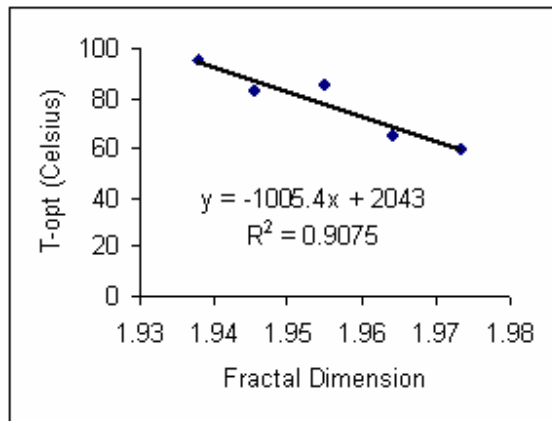


Fig. 3 Fractal dimension versus T-opt (optimal growth temperature) for the studied five archaea-eyryarchaeota thermophiles.

The high correlation ( $R^2 = 0.91$ ) of fractal dimension with T-opt suggests that the nucleotide fluctuation is not entirely random but may contain addition information influencing the morphological growth property. BLAST comparison showed that the sequences were about 75% identical. Computer simulation shows that random sequences would have an average of about 2 with a standard deviation about 0.015. The

observed fractal dimension values are outside the standard deviation interval consistent with the presence of selection pressure. The specificity of fractal dimension as a marker can be further tested by the incorporation of diverse species in the correlation analysis. The inclusion of two arachea-crenarchaeota thermophiles reduces the  $R^2$  value to 0.66 (Fig. 4).

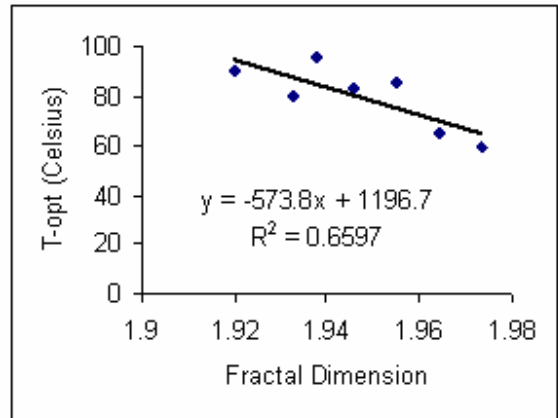


Fig. 4 Fractal dimension versus T-opt (optimal growth temperature) for the studied five archaea-eyryarchaeota thermophiles and two arachea-crenarchaeota thermophiles.

The mixing of arachea-crenarchaeota organisms, *aeropyrum pernix* and *Sulfolobus solfataricus*, with the archaea-eyryarchaeota organisms into a single dataset yields a correlation level similar to the dataset of bacterial radiation dosage and fractal dimension [1]. The relationship of the rec-A repair gene fractal dimension with radiation dosage of six organisms is displayed below for easy reference (Fig. 5).

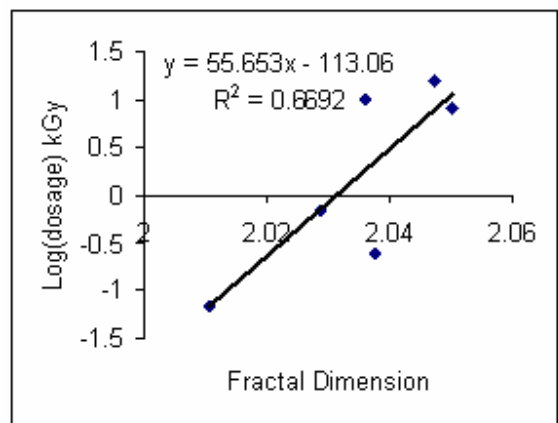


Fig. 5 Fractal dimension versus radiation dosage

The studied organisms were *kineococcus radiotolerans*, *deinococcus radiodurans*, *E. coli* K-12, *pseudomonas putida*, and *shewanella oneidensis*. The Genbank accession information of the rec-A repair gene sequences has been reported [14]. The low correlation of fractal dimension with phenotype as measured by T-opt and radiation dosage suggests that fractal dimension can serve as a sensitive marker

for closely related sequences, presumably closely related organisms. The fractal dimension positive correlation with radiation dosage can be interpreted as the necessary increase of information capacity for handling multitasking sequence repair under radiation. On the other hand the 16S rRNA has a very specialized function as a ribosome part and high temperature could have selected low information capacity sequence.

The further inclusion of two bacterial thermophiles reduces the  $R^2$  value to 0.50 (Fig. 6). Note that the Adjusted- $R^2$  is 0.33 compared to 0.82 for the five archaea-eyryarchaeota thermophiles complied in Fig. 3.

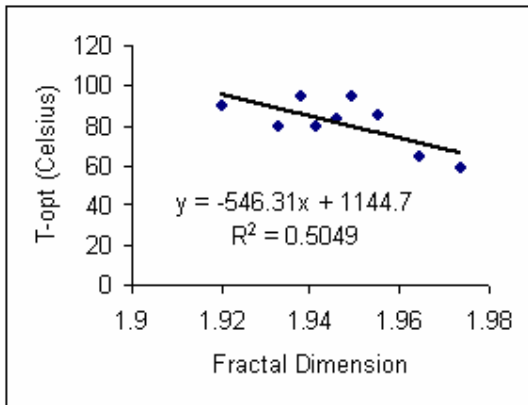


Fig. 6 Fractal dimension versus T-opt (optimal growth temperature) for the studied five archaea-eyryarchaeota thermophiles, two archaeae-crenarchaeota thermophiles, and two bacterial thermophiles.

The resulting poor correlation suggests that the added organisms are very different from the arachea, and is consistent with the bacterial nature of the added *thermotoga maritime* and *aquifex aeolicus* thermophiles in the regression analysis.

**B. Sequence GC content and Shannon Entropy Analysis**

The fractal dimension is correlated to the sequence GC content with an  $R^2$  value of 0.89 for the five archaea-eyryarchaeota thermophiles (Fig. 7).

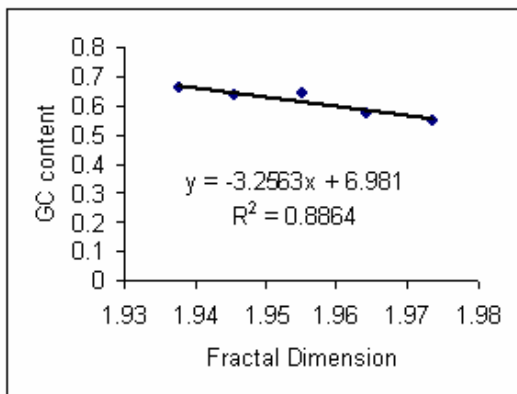


Fig. 7 Fractal dimension versus sequence GC content for the studied five archaea-eyryarchaeota thermophiles

In contrast to our data, computer simulation of purely random sequences shows little correlation of fractal dimension with GC content ( $R^2 \sim 0.1$  for GC content of about 65%). When taking into account of entire dataset of nine organisms, the correlation  $R^2$  value drops to 0.74 for the entire dataset of nine organisms (Fig. 8).

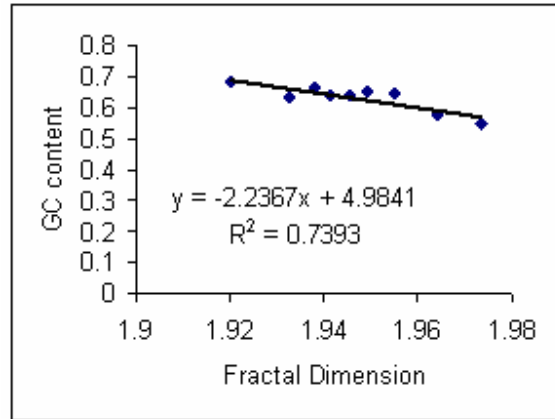


Fig. 8 Fractal dimension versus sequence GC content for the studied nine thermophiles

The highest correlation was founded to be between the fractal dimension and di-nucleotide Shannon entropy (Fig. 9). There are 16 di-nucleotide pairs and the probability of each di-nucleotide pair can be calculated from the sequence histogram. The Shannon entropy can be calculated using  $p \cdot \ln(p)$ .

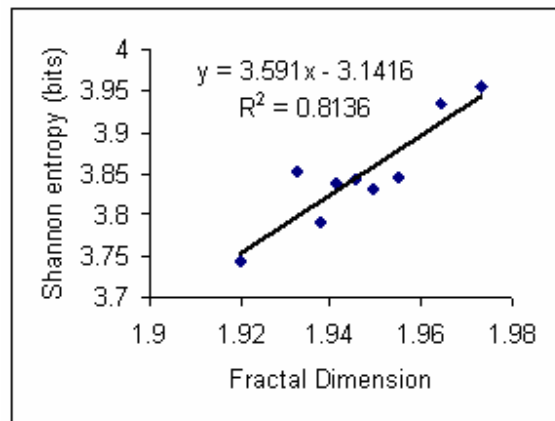


Fig. 9 Fractal dimension versus sequence di-nucleotide Shannon entropy for the studied nine thermophiles.

However, the di-nucleotide Shannon entropy was observed to correlate with optimal growth temperature with an  $R^2$  of 0.8 (Fig. 10). Furthermore, sequence GC content was observed to correlate with optimal growth temperature with an  $R^2$  of 0.88 for the entire set of 9 thermophiles (Fig. 11). Thus the correlation of T-opt with either di-nucleotide Shannon entropy or sequence GC content lacks species specificity. The relationship of GC content with heat capacity change during DNA folding has been reported [15, 16]. The stability of the

GC bond is important for many of our observed correlations, since having GC content much higher than 50% lowers both Shannon entropy and fractal dimension. Therefore fractal dimension remains as the only sensitive marker in this project for studying species specificity.

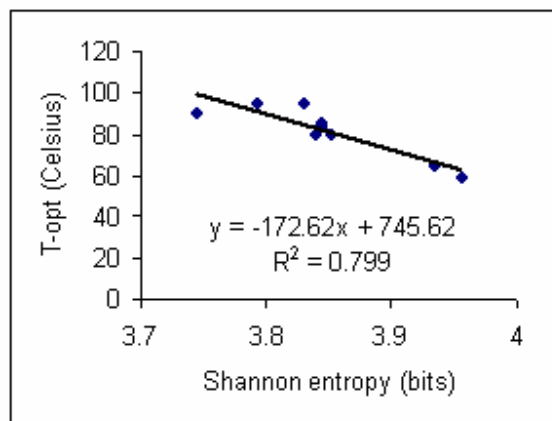


Fig. 10 Di-nucleotide Shannon entropy versus T-opt for the studied nine thermophiles

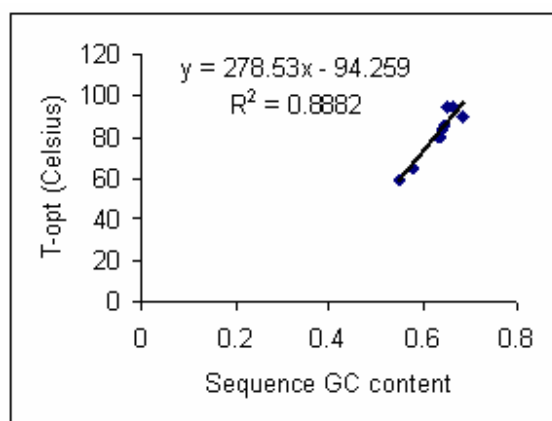


Fig. 11 Sequence GC content versus T-opt for the studied nine thermophiles

#### IV. CONCLUSION

Our results show that the fractal dimension of the 16S rRNA gene has a correlation with the optimal growth temperature using nine thermophiles. Sequences in closely related species show high correlation near 0.9 (Adjusted-R<sup>2</sup> 0.8) at the phylum level for the studied archaea-euryarchaeota organisms. The Adjusted-R<sup>2</sup> correlation was observed to degrade to 0.33 when using all thermophiles studied, including bacteria. The high correlations of either sequence GC content or di-nucleotide with the optimal growth temperature at the domain level limit their application as species specific markers. The study of sequence nucleotide fluctuation showed that fractal dimension analysis can be used to support the presence of selection pressure. It is postulated by us that fractal dimension analysis is a sensitive tool for studying the relationship between genotype and phenotype

among closely related sequences, possibly at the phylum level for the archaea domain.

#### ACKNOWLEDGMENT

The project was partially supported by several CUNY PSC and Collaborative grants. A.F., N.G., and R. Sullivan received partial support from CUNY New Faculty Programs. E.C. thanks the hospitality of QCC. We thank the research groups for posting their gene data in the public domain.

#### REFERENCES

- [1] Todd Holden, R. Subramaniam, R. Sullivan, E. Cheung, C. Schneider, G. Tremberger, Jr., A. Flamholz, D. H. Lieberman, and T. D. Cheung, "ATCG nucleotide fluctuation of *Deinococcus radiodurans* radiation genes", Proc. SPIE 6694, 669417, 2007
- [2] N. N. Ojwa and J. A. Glazier, "The fractal structure of the mitochondrial genomes", Physica A, vol 311, pp221 – 230, 2002.
- [3] Z.G. Yu, A. Vo, Z.M. Gong and S.C. Long, "Fractals in DNA sequence analysis", Chinese Physics, vol 11, pp1313-1318, 2002.
- [4] H.D. Liu, Z.H. Liu, X. Sun, "Studies of Hurst Index for Different Regions of Genes", ICBBE 2007, pp238-240, 2007.
- [5] C.Y. Lee, "Mass Fractal Dimension of the Ribosome and Implication of its Dynamic Characteristics", Physical Review E, vol 73, 042901 (3 pages), 2006.
- [6] Pollard KS, Salama SR, Lambert N, Coppens S, Pedersen JS, et al., "An RNA gene expressed during cortical development evolved rapidly in humans". Nature 443, 167-172, 2006.
- [7] Pollard KS, Salama SR, King B, Kern AD, Dreszer T, et al., "Forces shaping the fastest evolving regions in the human genome", PLoS Genet 2(10): e168. DOI: 10.1371/journal.pgen.0020168, 2006
- [8] E.W. Weisstein, "Capacity Dimension." From MathWorld--A Wolfram Web Resource. <http://mathworld.wolfram.com/>
- [9] Huai-chun Wang & Donal A. Hickey, "Evidence for strong selective constraint acting on the nucleotide composition of 16S ribosomal RNA genes", Nucleic Acid Research, vol 30, 2501-2507, 2002.
- [10] W. Klonowski "From conformons to human brains: an informal overview of nonlinear dynamics and its applications in biomedicine". Nonlinear Biomed Phys. 2007 Jul 5; 1(1):5.
- [11] M.J. Berryman, A. Allison, and D. Abbott, "Mutual Information for examining correlations in DNA", Fluctuation & Noise Letters, vol 4, ppL237-L246, 2004.
- [12] T. Higuchi, "Approach to an irregular time series on the basis of fractal theory", Physica D, vol 31, 277-283, 1998.
- [13] Xinmin Yang, Haluk Beyenal, Gary Harkin, Zbigniew Lewandowski, "Quantifying biofilm structure using image analysis", Journal of Microbiological Methods, Vol 39, Pages 109-119, 2000
- [14] Todd Holden, G. Tremberger, Jr., P. Marchese, E. Cheung, R. Subramaniam, R. Sullivan, P. Schneider, A. Flamholz, D. Lieberman, & T. Cheung, "DNA sequence based comparative studies of between non-extremophile and extremophile organisms with implications in exobiology", SPIE Astrobiology Conference Proceedings, 7097-30., invited, in press, 2008.
- [15] Stoyan Milev, Alemayehu A. Gorfe, Andrey Karshikoff, Robert T. Clubb, Hans Rudolf Bosshard, and Ilian Jelesarov, "Energetics of Sequence-Specific Protein-DNA Association: Binding of Integrase Tn916 to Its Target DNA" Biochemistry vol 42, 3481-3491, 2003.
- [16] Stoyan Milev, Alemayehu A. Gorfe, Andrey Karshikoff, Robert T. Clubb, Hans Rudolf Bosshard, and Ilian Jelesarov, "Energetics of Sequence-Specific Protein-DNA Association: Conformational Stability of the DNA Binding Domain of Integrase Tn916 and Its Cognate DNA Duplex" Biochemistry vol 42, 3492-3502, 2003.