

# On Complexity Measures for Biological Sequences\*

Fei Nan and Donald Adjeroh

**Abstract.** In this work, we perform an empirical study of different published measures of complexity for general sequences, to determine their effectiveness in dealing with biological sequences. By effectiveness, we refer to how closely the given complexity measure is able to identify known biologically relevant relationships, such as closeness on a phylogenetic tree. In particular, we study three complexity measures, namely, the traditional Shannon's entropy, linguistic complexity, and T-complexity. For each complexity measure, we construct the complexity profile for each sequence in our test set, and based on the profiles we compare the sequences using different performance measures based on: (i) the information theoretic divergence measure of relative entropy; (ii) apparent periodicity in the complexity profile; and (iii) correct phylogeny. The preliminary results show that the T-complexity was the least effective in identifying previously established known associations between the sequences in our test set. Shannon's entropy and linguistic-complexity provided better results, with Shannon's entropy having an upper hand.

## 1. Introduction

The complexity of an organism has a direct manifestation in the general organization of its genomic structures. Given the primary genomic sequence for an organism, we can make certain predictions about the organism based on the randomness of the sequence, or the level of difficulty in predicting or compressing the sequence. Thus sequence complexity plays an important role in various application areas, such as in biological sequence compression for compact storage of the sequence, construction of phylogenetic trees, comparative genomics, studies of genomic evolution, etc. The genomic complexity has also been linked to the amount of information an organism stores about its environment [Adami2000c].

Various measures of complexity have been proposed in the literature. For instance, Allison et al [Allison2000sed] proposed a statistical method that considers both forward and reverse repeats in the sequence, including complementary repeats. The notion of physical complexity was proposed by Adami et al

[Adami2000c, Adami2000oc], where complexity was viewed as the difference between the maximal entropy of an ensemble and the actual entropy, given a specific environmental condition. Methods based on compositional complexity of a sequence were proposed in [Wan2000w, Roman-Roldan98bo]. Gusev et al [Gusev99nc] proposed complexity measures for genetics sequences, based on a modification of the general sequence complexity measure described by Lempel and Ziv [Lempel76z]. More recently, in [Taft2003m], a simple measure of genome complexity, based on the ratio of non-coding DNA to the total DNA was proposed, and used to argue that the amount of non-coding DNA may have a positive correlation with the complexity of the organism. See also [Adami2002] for an interesting article on the general notion of complexity.

In this work, we study different published measures of complexity for general sequences, to determine their effectiveness in dealing with biological sequences. In particular, we study three complexity measures: the traditional Shannon's entropy [Cover91t], linguistic complexity [Troyanskaya2002aklb], and T-complexity [Ebeling2001st]. For each complexity measure, we construct the complexity profile for each sequence in our test set, and based on the profiles we compare the sequences using different measures, based on the Kullback-Leibler divergence [Cover91t], and the apparent periodicity in the complexity profile. We measure their effectiveness based on the extent to which they can distinguish between (or relate) different known sequences, and the extent to which their complexity ranking of the sequences compares with the complexity ranking determined by specific ground-truths. In the next section, we give precise definitions for the complexity measures we used. Section 3 describes the methods we used to compare the complexity measures. Results are presented in section 4.

## 2. Complexity Measures

### 2.1 Complexity Measures

The three complexity measures are described below.

#### 2.1.1 Shannon's Entropy

\* Authors are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506. email: [fein, don]@csee.wvu.edu. This work was partially supported by a DOE CAREER Grant, No.: DE-FG02-02ER25541.

This is the usual entropy, which is defined based on the probability of occurrence of the symbols. The entropy of a sequence  $S$  is defined as:

$$H(S) = -\sum_{i \in \Sigma} p(s_i) \log p(s_i),$$

where  $p(s_i)$  is the probability of the  $i$ -th symbol in the alphabet,  $\Sigma$ . For Shannon's entropy to approach the true entropy of the sequence, we have to consider the higher-order dependencies in the sequence. Dependencies can be considered in terms of some higher-order entropy – which could be different, depending on if we consider a simple extension of the source (i.e.  $n$ -blocks of symbols a time), or if we consider possible hidden Markov relationships in the sequence [Cover91t]. The entropy of the  $n$ -th extension of the source is given by:

$$H(S^n) = -\sum_{i \in \Sigma^n} p(s_i^n) \log p(s_i^n),$$

where  $|\Sigma|^n$  is now the alphabet of the extended source,  $p(s_i^n)$  is the probability of the  $i$ -th symbol in the new alphabet. For zero-memory sources,  $p(s_i^n) = p(s_1 s_2 \dots s_n) = p(s_1) p(s_2) \dots p(s_n)$  where  $s_1 s_2 \dots s_n$  is an  $n$ -block of symbols. With the entropy from the extended source, the entropy of the original source is given as:

$$h = \lim_{n \rightarrow \infty} \frac{1}{n} H(S^n)$$

The problem is the amount of data and time that may be required for reliable computation of the higher-order entropies. The discussion and results in this work are based on the first order entropy.

## 2.12 T-Complexity

The T-complexity[Eberling2001st] is similar in spirit to the production complexity for finite sequences described by Lempel and Ziv[Lempel76z]. Essentially, given a sequence  $S$ , the production complexity measures how difficult it will be to generate  $S$  from its symbol alphabet  $\Sigma$ . This in turn depends on the size of the vocabulary that will be generated from  $S$ , based on a specific decomposition algorithm. The T-complexity, on the other hand, measures the effective number of T-augmentation steps that will be needed to generate the given sequence  $S$ , from its alphabet  $\Sigma$ .

The T-complexity is computed as follows [Eberling2001st]: First parse  $S$  into its constituent patterns  $s_i$ , each with a respective copy exponent  $c_i$ , such that.

$$S = s_n^{c_n} s_{n-1}^{c_{n-1}} \dots s_i^{c_i} \dots s_1^{c_1} \sigma_0,$$

where  $s_i \in \Sigma^+$ ,  $\sigma_0 \in \Sigma$ ,  $i = 1, 2, 3, \dots, n$ ,  $c_i = 1, 2, 3, \dots$

Here,  $\Sigma^+ = \Sigma^* \setminus \Lambda$ , denotes the set of all finite strings from,  $\Sigma$  excluding the empty string,  $\Lambda$ . The T-complexity for  $S$  is then defined in terms of the copy exponents:

$$C_T(S) = \sum_{i=1}^n \log_e (c_i + 1)$$

The constituent patterns  $s_i$  are required to meet a specific constraint:

$$s_i = s_{i-1}^{m_{i,j-1}} s_{i-2}^{m_{i,j-2}} \dots s_{i-j}^{m_{i,j}} \dots s_1^{m_{i,1}} \sigma_i,$$

where  $\sigma_i \in \Sigma$ , and  $0 \leq m_{i,j} \leq c_j$ . In general, the T-complexity is minimal for a sequence of the form,  $S = \sigma^k$ , a  $k$ -length sequence of the same symbol. It is also easy to see that for  $k = |S|$ ,  $C_T(S) \geq \log_e(k)$ .

## 2.1.3 Linguistic Complexity

The linguistic complexity [Troyanskaya2002aklb] also seeks to exploit the size of the distinct vocabulary in determining the complexity of an input string, although in a different way. Given a sequence  $S$  of length  $k = |S|$ , with symbols from the alphabet  $\Sigma$ , the linguistic complexity (LC) for  $S$  is simply defined as the ratio of the number of distinct substrings in  $S$  to the maximum possible number of distinct substrings for a sequence of length  $k$ , using the same alphabet,  $\Sigma$ . The maximum possible number of distinct substrings for a sequence of length  $k$  is essentially the maximum vocabulary size, given by:

$$V_{\max}(k, |\Sigma|) = \sum_{v=1}^k \min\{|\Sigma|^v, k - v + 1\}.$$
 Thus,

$$LC(S) = \frac{\# \text{ distinct substrings in } S}{V_{\max}(|S|, |\Sigma|)}$$

The maximum LC of 1 occurs when all the possible substrings occur in the sequence. Again, for a given sequence length,  $k$ , the minimum value for LC occurs for sequences of the form,  $S = \sigma^k$ . We observe that this minimal value depends on the length of the sequence.

## 2.2 Complexity Profiles

An important issue in considering complexity for sequences is the problem of locality. While the usual complexity measures such as entropy provide one single value for a given sequence, independent of the sequence length, it is known that the complexity varies significantly over the sequence. More importantly, low complexity zones along the sequence are often symptomatic of areas of important biological significance. For measures such as linguistic complexity, the use of different window sizes is one way to capture possible local complexity variations in the sequence. We use the same windowing concept for the two other measures. That is, for a measure such as entropy, we compute the entropy using an overlapping window along the sequence. The result is a sequence complexity profile, which shows the variation of complexity along the length of the sequence. It is clear that the window

size will have a direct influence on the complexity profile. For our tests, we used three window sizes:  $w=50$ ,  $w=100$ , and  $w=200$ . Fig. 1 shows plots of the complexity profile for one sequence in our data set, using the three complexity measures ( $w=100$ ).

### 3. Performance Measures

#### 3.1 Ground Truth

To compare the performance of the three complexity measures, we used a set of seven gene sequences, from a recently published work [Chen2000kl]. The sequences are taken from three species as follows:

**Archaeobacteria:** *H. Butylicus* (H\_b); *Halobaculum gomorrense* (H\_g)

**Eubacteria:** *Aerococcus urina* (A\_u); *M. glauca* (M\_g); *Rhodopila globiformis* (R\_g)

**Eukaryotes:** *Urosporidium crescens* (U\_c); *Labyrinthula sp. Nakagiri* (L\_n)

The sequences are available in GENBANK<sup>2</sup>. Items in brackets correspond to labels as used in this report.

As our ground truth, we applied different compression algorithms on each of the sequences, and then use the average compression performance (i.e. average bits per symbol as reported by the different algorithms) as a measure of complexity. We assume that the most complex sequence should result in the least compression, or largest value in terms of bits per symbol. Table I shows the results, and the corresponding first order entropy for each sequence. Further, for comparative results, we assume the phylogenetic tree reported in [Chen2000kl] as a ground truth, in terms of the closeness between the sequences. Using this tree, the sequences are grouped as follows: **G1:** H\_b, H\_g; **G2:** A\_u, M\_g, R\_g; **G3:** L\_n, U\_c. Here, we compare the complexity measures in terms of how well they could group similar sequences together.

#### 3.2 Direct Measurement

By direct measurement, we refer to measurements that can be made directly on the complexity profile for a given sequence, without regard to the other sequences. Thus, the results here can be used to judge how well the complexity measures could rank the sequences, when compared to known complexity rankings.

##### 3.2.1 Average of Complexity Profile

For a given complexity profile, we compute the mean  $\mu$ , and standard deviation  $\sigma$ , to obtain a weighted sum:

$$f(\mu, \sigma) = w_\mu \mu + w_\sigma \sigma, \quad \text{where,} \quad w_\mu + w_\sigma = 1, \quad \text{and}$$

$w_\mu, w_\sigma$  are weights. We use  $w_\mu = 0.4, w_\sigma = 0.6$  in our tests. To avoid bias,  $\mu$  and  $\sigma$  are each normalized to the range [0 1] before using them in the calculations.

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/>

#### 3.2.2. Apparent Periodicity

The apparent periodicity measures how periodic the complexity profile is. Clearly, sequences that are complicated are expected to exhibit a less periodic complexity profile. To assess the periodicity, we consider the sequence as a one-dimensional data, and compute the apparent periodicity using the mean and standard deviation, as follows:

Let  $x_i$  denote the complexity value at the  $i$ -th point along the sequence.

Define  $f_j(x_i) = x_i - (\mu + j\sigma)$ , where  $j = -2, -1, 0, 1, 2$ .

$$g(x_i) = \begin{cases} 1; & x_i = x_1 \\ 1; & \text{sign}(f(x_i)) \neq \text{sign}(f(x_v)) \wedge |f(x_v)| \geq \Delta \\ 0; & \text{otherwise} \end{cases}$$

where  $v = \arg \min_q \{x_i - x_q \text{ s.t. } g(x_i) = 1\}$ , and  $\Delta$  is a

small number. Let  $P = \sum_{i=1}^k g(x_i)$ . For each of the values of  $j$  above, we compute the corresponding  $P$  (namely,  $P_0, P_1, P_{-1}, P_2, P_{-2}$ ), and combine these to

determine the overall periodicity:  $P_{ave} = \sum_{j=1}^5 w_j P_j$ .

We then obtain the final periodicity by considering the length of this sequence  $k$ :

$$\text{Apparent periodicity, } \rho = \frac{1}{k} P_{ave}$$

$$\text{Period length: } L_\rho = 1/\rho$$

We use  $w_0 = 0.5; w_1 = w_{-1} = 0.15; w_2 = w_{-2} = 0.1$  in our tests.

#### 3.3 Comparative Measurement

##### 3.3.1. Relative Entropy

For comparative measurement, we use the relative entropy (also called the Kullback-Leibler distance) between the complexity profiles. First, we generate normalized histograms from the complexity profiles, and based on the normalized histogram counts, we compute the relative entropy between pairs of profiles. For two given probability distributions,  $P = \{p_1, p_2, \dots, p_{|\Sigma|}\}$  and

$Q = \{q_1, q_2, \dots, q_{|\Sigma|}\}$ , the relative entropy is defined as

$$\text{follows [Cover91t]: } D(P \| Q) = \sum_{i=1}^{|\Sigma|} p_i \log \frac{p_i}{q_i}.$$

Since  $D(\cdot \| \cdot)$  is not symmetric, we compute both  $D(P \| Q)$  and  $D(Q \| P)$ , and take the average. Sequences that are similar should have small relative entropy, while the relative entropy should be large for profiles that are less similar.

##### 3.3.2. Phylogeny and Classification

For a further comparative measure, we use the mean and standard deviation of the profiles to perform a

classification of the sequences, to see how close the grouping will be with that of the ground truth phylogenetic tree. Here, we use  $\mu$  and  $\sigma$  as two dimensions for the classification, and use the scatter plots to evaluate the closeness (or otherwise) of the sequences.

#### 4. Results

The results are shown in Figs.1 and 2, and Tables I – IV. (See last page for Figs. 1 and 2 and Table I). With the direct measurements, none of the tested complexity measures was consistent in matching the complexity ranking produced by the ground truth (empirical compression ratio). It may also be observed that the ranking due to compression performance did not necessarily agree with that due to first order entropy on the entire sequence. However, the direct measurements provide an idea towards comparative analysis of the complexity. See for example Shannon complexity in Table I.

**Table II: Average Complexity: mean and standard deviation ( $w=100$ )**

	Entropy	SC	LC	TC	Ave. Compr.
<b>M_g</b>	1.9807	0.4070	0.4360	0.1463	2.7136
<b>A_u</b>	1.9852	0.4250	0.4000	0.1562	2.6951
<b>H_g</b>	1.9728	0.8352	1.0000	0.8511	2.6938
<b>R_g</b>	1.9617	0.3582	0.3677	0.0672	2.6694
<b>L_n</b>	1.9857	0.4649	0.4915	0.0661	2.6605
<b>H_b</b>	1.8861	0.5045	0.6000	0.6986	2.6543
<b>U_c</b>	1.9794	0.4636	0.5777	0.1509	2.6174

**Table III: Periodicity results (period length)**

	Shannon Complexity			Linguistic Complexity			T-Complexity		
	w=50	100	200	50	100	200	50	100	200
<b>A_u</b>	37.10	79.20	197.43	35.30	137.51	691.00	4.41	5.14	5.05
<b>H_b</b>	56.13	103.40	154.79	47.12	113.01	370.14	6.24	7.80	8.68
<b>H_g</b>	97.09	142.28	185.42	37.57	147.38	487.12	5.58	6.83	7.47
<b>L_n</b>	42.16	73.51	168.85	42.16	117.46	810.50	4.57	5.03	5.01
<b>M_g</b>	26.18	69.38	117.97	43.12	106.27	743.89	5.10	4.77	5.53
<b>R_g</b>	29.10	48.11	106.55	33.76	110.15	418.57	4.47	4.69	5.14
<b>U_c</b>	34.71	88.33	168.89	40.26	116.92	688.30	4.85	4.80	4.78

**Table IV: Detailed periodicity results (Shannon complexity,  $w=100$ ).**

	P <sub>0</sub>	P <sub>1</sub>	P <sub>1</sub>	P <sub>2</sub>	P <sub>2</sub>	Pavg	Periodicity	Period Length
<b>A_u</b>	24	10	21	1	7	17.45	0.0126	79.1977
<b>H_b</b>	14	25	8	1	3	12.35	0.0097	103.4008
<b>H_g</b>	17	1	7	1	3	10.10	0.0070	142.2772
<b>L_n</b>	31	14	19	1	15	22.05	0.0136	73.5147
<b>M_g</b>	30	11	11	1	9	19.30	0.0144	69.3782
<b>R_g</b>	41	20	37	3	11	30.45	0.0208	48.1117
<b>U_c</b>	26	16	27	1	11	20.65	0.0113	88.3293

Results for the comparative measures are shown in Fig. 2. The scatter plots clearly shown that some of the complexity measures were able to correctly group

similar sequences together, in accord with the ground truth results of phylogeny.

#### 5. Conclusion

We have studied the performance of three measures of complexity with respect to biological sequences, based on direct measurements on the complexity profiles, and on relative comparisons with other profiles. The results for direct measurements were inclusive, as none of the measures was consistent in reproducing the known ranking of the test sequences, as produced by a various compression systems. For comparative measurements, Shannon's entropy produced the best result, followed by the linguistic complexity.

#### References:

[Adami2000c] Adami C., and Cerf N.J., "Physical complexity of symbolic sequences", *Physica D* 137, 62-69, 2000.

[Adami2000coc] Adami C., Ofria C., and Collier T.C., "Evolution of biological complexity", *Proc. Nat. Acad. Sci (USA)* 97, 4463, 2000

[Adami2002] Adami C., "What is complexity", *Bioessays*, 24(12):1085-94, 2002.

[Allisson2000] Allison L, Stern L, Edgoose T and Dix TI, "Sequence complexity for biological sequence analysis", *Computers & Chemistry*, 24(1), 3-55, 2000.

[Chen2000kl] Chen X, Kwong S, and Li M, "A compression algorithm for DNA sequences and its applications in genome comparison", *Proc., 4<sup>th</sup> Annual Conference on Research in Computational Molecular Biology*, Tokyo Japan, pp. 107, 2000.

[Cover91t] Cover T.M. and Thomas J. A., *Elements of Information Theory*, Wiley, 1991

[Eberling2001st] Ebeling W, Steuer R, and Titchener MR, "Partition-based entropies of deterministic and stochastic maps", *Stochastics and Dynamics*, 1, 1 1 – 17, 2001.

[Gusev99nc] Gusev VD, Nemytikova LA, Chuzhanova NA, "On the complexity measures of genetic sequences", *Bioinformatics*, 15(12):994-999, 1999.

[Lempel76z] Lempel, A., and Ziv, J., "On the complexity of finite sequences", *IEEE Transactions on Information Theory*, 22, 21-27, 1976

[Taft2003m] Taft RJ, and Mattick JS, "Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences", *Genome Biology*, 5:P1, 2003 (deposited research article).

[Troyanskaya2002aklb] Troyanskaya OG, Arbell O, Koren Y, Landau GM, and Bolshoy A, "Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating

linguistic complexity”, *Bioinformatics*, 18 (5), 679-688, 2002;

[Wan2000w] Wan H, Wootton JC., “A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode

less complex proteins”, *Comput Chem.*, 24(1):71-94, 2000.

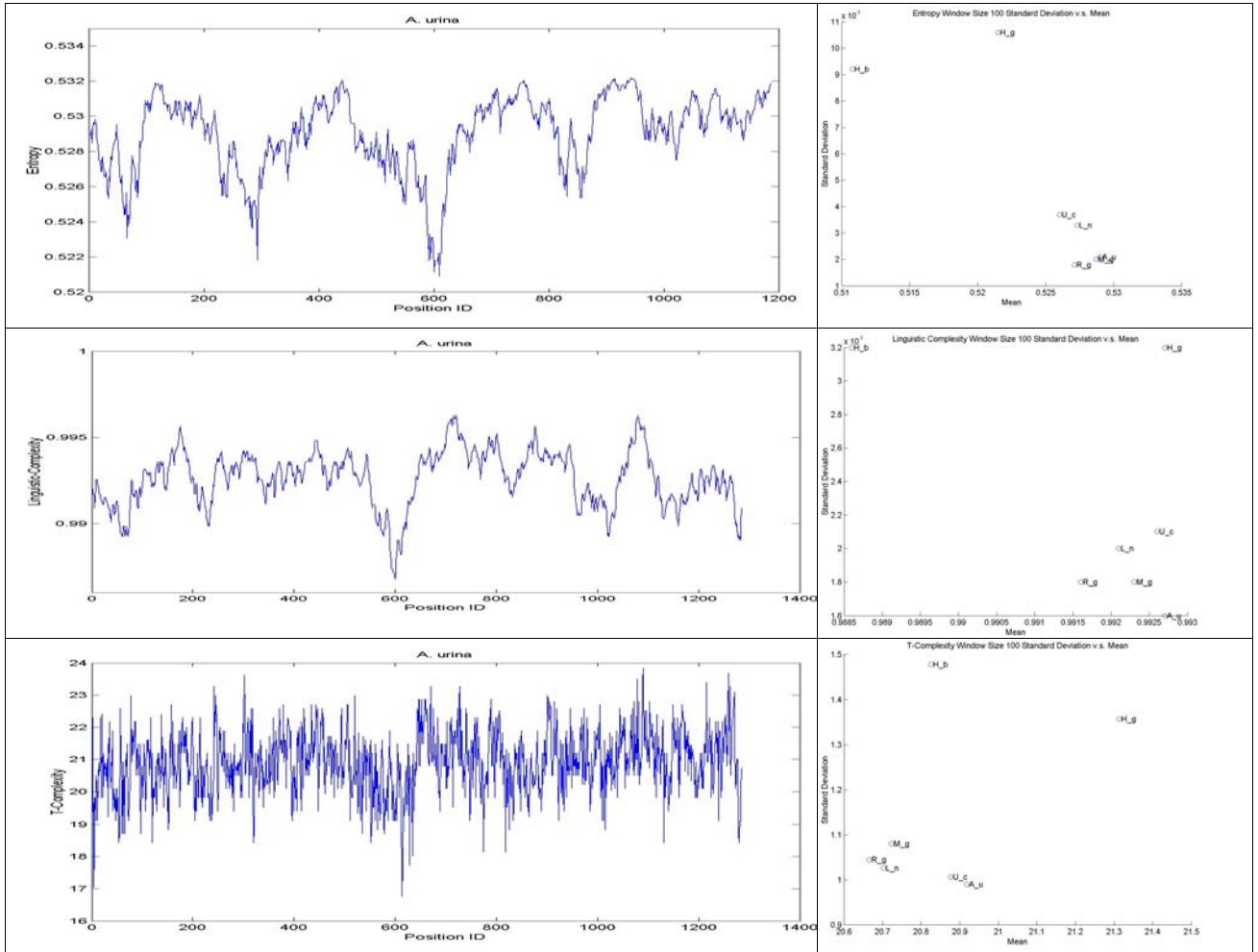


Fig1 (col. 1): Complexity profile for a sample sequence in the test set.

Fig.2 (col. 2) : scatter plots for SC, TC, LC.

Table I: Ground Truth - Compression Results (in bits/symbol) for the sequences.

Entries are ordered according to the Average compression result (column 4).

	Length	Entropy	Ave. Result	Detailed Compression Results						
				Gen Compress	DNA Compress	BWT	PPM	WinZip	GZip	Compress
<b>M_g</b>	1339	1.981	2.71	2.025	2.10	2.74	2.74	3.48	2.93	2.98
<b>A_u</b>	1382	1.985	2.70	2.026	2.10	2.72	2.71	3.45	2.92	2.94
<b>H_g</b>	1437	1.973	2.69	2.026	2.09	2.69	2.73	3.48	2.95	2.89
<b>R_g</b>	1465	1.962	2.67	2.026	2.07	2.67	2.69	3.45	2.92	2.86
<b>L_n</b>	1621	1.986	2.66	2.023	2.09	2.65	2.68	3.38	2.91	2.89
<b>H_b</b>	1277	1.886	2.65	2.030	1.98	2.66	2.64	3.52	2.93	2.82
<b>U_c</b>	1824	1.979	2.62	2.022	2.07	2.59	2.64	3.29	2.87	2.84