EUROPEAN TRANSACTIONS ON TELECOMMUNICATIONS
Eur. Trans. Telecomms. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/ett.1201

On genomic coding theory†

Zaher Dawy[1]*, Pavol Hanus[2], Johanna Weindl[2], Janis Dingel[2] and Faruck Morcos[3]

[1]*Electrical and Computer Engineering Department, American University of Beirut, Lebanon*
[2]*Institute for Communications Engineering, Munich University of Technology, Germany*
[3]*Department of Computer Science and Engineering, Notre Dame University, IN, USA*

SUMMARY

This paper gives a brief overview of several applications from the emerging interdisciplinary field of genomic coding theory that aims at applying concepts and techniques from the field of coding theory to problems from the field of molecular biology. This is motivated by the high precision and robustness found in genomic processes in addition to the increase in the availability of genomic data for a wide range of species. The considered applications include source coding for DNA classification, channel coding for modelling gene expression with emphasis on the process of translation, existence of error correcting codes in the DNA and channel coding structure in the genetic code. Example results are presented that demonstrate the relevance of the proposed approaches and open questions are formulated to suggest future research work.

1. INTRODUCTION

Motivated by the redundant structure of the genetic code, the existence of large evolutionary conserved non-coding regions among species, and the existence of special sequences in coding regions, several researchers are trying to apply coding theory models to understand the structure of the DNA and the operation of various genetic processes.

Yockey [1] proposed one of the first models for gene expression using encoding/decoding concepts from communication theory. Liebovitch *et al*. [2] developed the first efficient method to scan through DNA sequences to determine whether some linear block code structure is present. Years later, Rosen [3] developed a method for the detection of linear block codes that accounts for possible insertions and deletions in the DNA sequences. However, neither work was able to support the existence of such simple error correcting codes in the DNA. Battail [4] argued about the existence of nested error correcting codes in the DNA supported by several biological observations such as the size of the human genome being far larger than the size needed to specify every characteristic of any given individual. On other fronts, Mac Donaill [5] proposed a parity check code interpretation of nucleotide composition, and May *et al*. [6] proposed the use of block and convolutional codes to model the process of translation initiation in prokaryotic organisms.

This paper is organised as follows. Section 2 presents some biological background on gene expression in addition to some analogies that motivate this work. Section 3 presents recent research contributions and open problems in the field of genomic coding theory. Finally, conclusions are drawn in Section 4.

2. WHY CODING THEORY?

2.1. From DNA to proteins

Gene expression is the process through which information contained in the DNA is transformed into proteins. Gene

* Correspondence to: Zaher Dawy, Electrical and Computer Engineering Department, American University of Beirut, Riad El Solh, Beirut 1107 2020, Lebanon. E-mail: zaher.dawy@aub.edu.lb
† Dedicated to Prof. Dr-Ing. Dr-Ing. E. h. Joachim Hagenauer on the occasion of his 65th birthday.

Accepted 10 January 2007

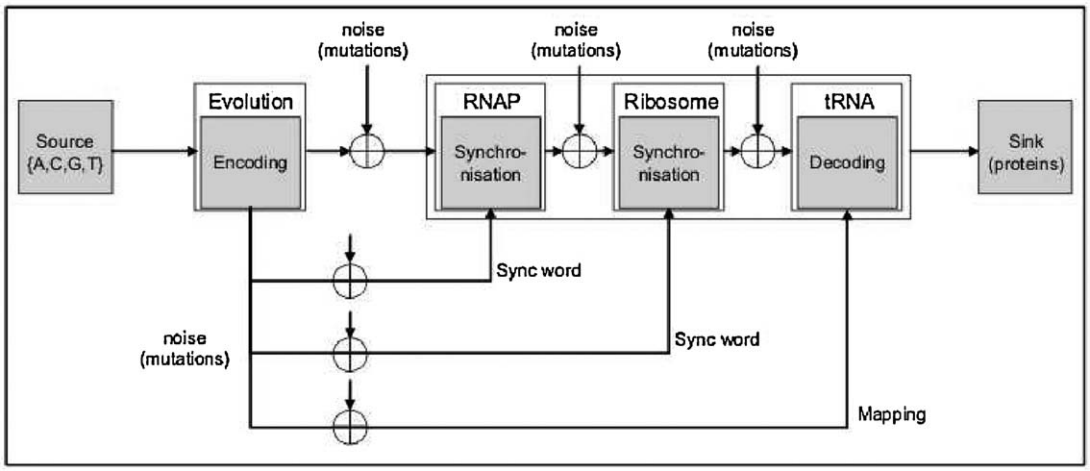

Figure 1. Communication theory model for gene expression.

expression is composed of two main steps: transcription and translation. In transcription, the double stranded DNA molecule is used to synthesise a new single stranded molecule called messenger RNA (mRNA). The RNA polymerase binds to a specific region in the DNA in order to separate the two strands. Once they are separated, one of the strands serves as a template for the creation of the mRNA. The resulting mRNA is consequently spliced to remove the introns (non-coding regions) which results in a sequence of pure exons called mature mRNA. The mature mRNA travels in the cell until the ribosome binds to it at a specific region in order to start the process of translation. Once the ribosome binds properly (translation initiation), it starts processing triplets of bases (also called codons) of the mature mRNA to produce amino acids. The ribosome serves as a platform for the transfer RNA (tRNA) molecule which holds the amino acids. A tRNA molecule connects using its anti-codon end with codons found in the mature mRNA until a sequence of amino acids is chained. The formed chain then folds to finally produce a protein. Note that the gene expression processes differ between eukaryotes and prokaryotes. Prokaryotes are simple organisms that do not have a nucleus such as bacteria (e.g. E. Coli) whereas eukaryotes are organisms that have their DNA in the nucleus (e.g. humans).

2.2. Analogies and modelling

There are several analogies between data transmission in communication systems and DNA processing in gene expression. The DNA can be modelled as an encoded information source that is decoded (processed) in several steps to produce proteins. During these decoding steps, the processed DNA is subjected to genetic noise which results in several types of mutations. Transcription initiation corresponds to a process of frame synchronisation where the RNA polymerase detects the promoter sequences (biological sync words). Translation initiation also corresponds to a process of frame synchronisation to detect the translation initiation signals (e.g. for prokaryotes this includes the Shine-Dalgarno (SD) sequence and the start codon). This is followed by a decoding process to map codons to amino acids. Figure 1 shows a model for gene expression based on building blocks from communication theory. In this model, we assume that mutations can also occur in the involved proteins, that is RNA polymerase, ribosome and tRNA. Other similar models for gene expression are summarised in Reference [6].

On a larger scale, evolution can be modelled as a single input multiple output (SIMO) antenna system. Given an evolutionary scenario of multiple species that evolved from a common ancestor, the ancestor can be modelled as the transmitter and its sequence of bases as the output of the information source. This information is transmitted over the branches of the evolutionary tree (phylogenetic tree), where the leaves of the tree correspond to the receive antennas of the SIMO system. The antennas receive the sequences that one can observe in different species with errors (mutations) occurring during the transmission process. These analogies motivate the use of coding theory for genetics.

DOI: 10.1002/ett

3. GENOMIC CODING THEORY

3.1. Source coding for DNA classification

Special source coding algorithms have been developed to compress genomic sequences by taking into account their structural properties, for example, DNACompress [7]. In Reference [8], we make use of source coding algorithms to approximate a mutual information-based distance measure for the classification of DNA sequences. The mutual information $I(S_i; S_j)$ between two given sources S_i and S_j can be transformed into a bounded distance measure through normalisation by their maximum possible mutual information. The resulting measure can then be expressed as

$$d(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\min(H(S_i), H(S_j))} \tag{1}$$

where $H(S_i)$ is the entropy of the source S_i. The compression achieved on a sequence generated by a given source is used to approximate its entropy. Moreover, the compression achieved on the concatenation of two sequences generated by the two compared sources is used to approximate their conditional entropy [9].

To demonstrate the performance of the given distance measure, we present results for differentiating between non-genic regions (ng), exons (ex) and introns (in). As content sequences, the first 50 000 nucleotides (50 kb) of each type are selected from human chromosome 19 (c19). As unknown sequences, groups of nucleotides of different sizes of each type are selected from human chromosome 1 (c1). For each unknown sequence i, the distance $d(S_i^{\mathcal{U}}, S_j^{\mathcal{C}})$ to every content sequence j is calculated. The unknown sequence is then classified as the type of the content sequence yielding the smallest distance. Using DNACompress with the given distance measure, all unknown sequences were correctly recognised (see Table 1).

Table 1. DNA classification using source coding.

$S_i^{\mathcal{U}} \backslash S_j^{\mathcal{C}}$	c19ng-50kb	c19in-50kb	c19ex-50kb
c1ng-300kb	0.041-best	0.842	1.025
c1ng-13kb	0.651-best	1.013	1.009
c1in-300kb	0.933	0.585-best	1.014
c1in-13kb	1.000	0.052-best	1.068
c1ex-300kb	1.017	1.006	0.963-best
c1ex-13kb	0.985	0.944	0.830-best

3.2. Channel coding for modelling gene expression

Concepts from coding theory can be used to develop biologically-motivated models for the processes of transcription and translation. This allows the analysis of various interactions that take place in gene expression using efficient computer simulations saving laboratory sources and time spent on experiments. Moreover, it might help in discovering new facts that allow further understanding of these processes. The work of May [6] established the first concrete ideas for modelling gene expression interactions based on algorithms inspired from coding theory.

The process of translation in prokaryotes is triggered by the detection of a biological sync word known as the SD sequence which is located around 10 bases before the translation start codon AUG. It has been stated that the last 13 bases of the 16SrRNA subunit of the ribosome, that binds to the mRNA, play an important role in the detection of the SD sequence [10]. In Reference [11], we model this detection/recognition system by designing a one dimensional codebook consisting of the nine sub-sequences with length $N = 5$ of the last 13 bases of the 16SrRNA molecule, that is, we obtain nine codewords $\mathbf{c}_i = [s_i, \ldots, s_{i+4}], i \in [1; 9]$ where $\mathbf{s} = [s_1, \ldots, s_{13}]$ denotes the sequence of the last 13 bases. A sliding window is applied on a given noisy mRNA sequence to select sub-sequences of length N and compare them with all codewords in the proposed codebook. The codeword that results in the minimum Hamming distance is selected and the obtained minimum metric value is recorded. For the analysis, we apply the algorithm to 1500 *E.coli* translated sequences annotated in the NCBI database [12]. In addition, we apply the algorithm on a set of *E.coli* sequences that contain a start codon but are not translated. Average results for both classes of sequences are plotted in Figure 2.

The x-axis represents the position in the aligned sequences. It can be seen that the proposed algorithm is able to identify the SD (peak at position 40) and the start codon (peak at position 51) in the translated sequences and, thus, is capable of differentiating between translated and untranslated sequences. Moreover, these results support the arguments for the importance of the 16SrRNA in the translation process.

To make biological use of the developed algorithm, it was applied to test the effect of single point mutations in the ribosome on protein synthesis. To do this, we have introduced point mutations in all positions of the last 13 bases of the 16SrRNA and executed the algorithm on the *E.coli* data set. The obtained results are summarised in Table 2 by quantising into five levels the influence of these

DOI: 10.1002/ett

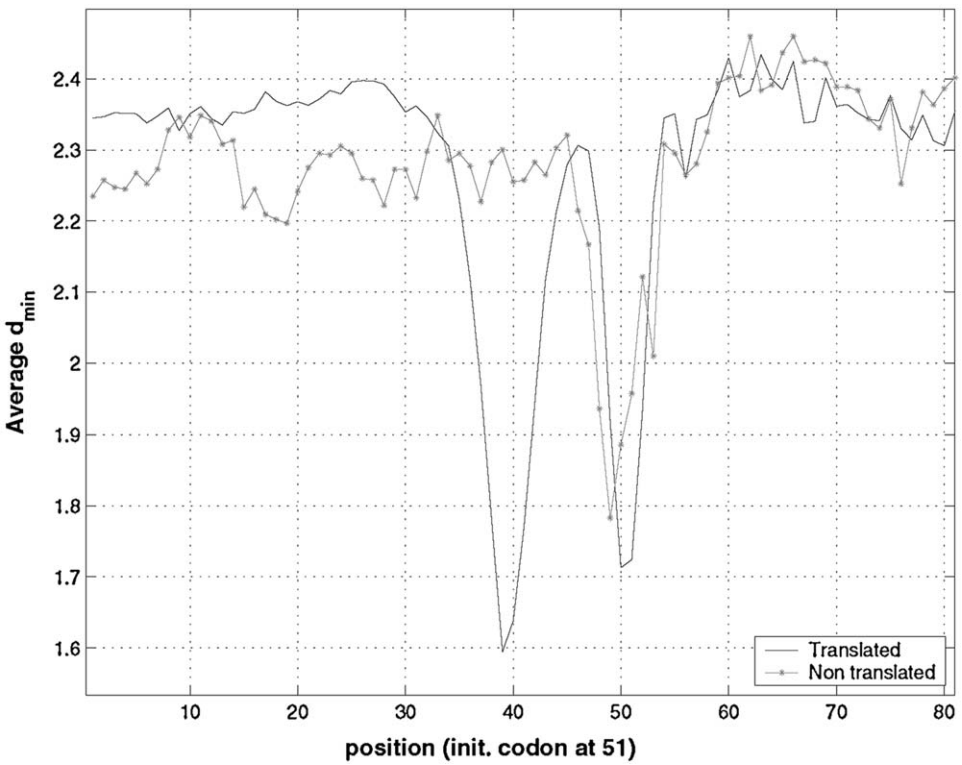

Figure 2. Detection of translation signals.

Table 2. Effect of mutations in the 16SrRNA on translation.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
SD	–	–	–	↓	⇓	⇓	↓	–	–	–	–	–	–
Start	–	–	⇓	↑	↓	↓	↓	–	⇓	↓	↓	↓	↑

mutations on detecting the SD and start codon signals. The levels are: – represents no influence, ⇓ a strong negative influence, ↓ a weak influence, ↑ a weak positive influence and ⇑ a strong positive influence.

For example, results show that a mutation in position 5 has a strong negative influence on the recognition of the SD signal, whereas a mutation in position 8 has no influence since perhaps its role is just to introduce spacing at the moment of decoding the mRNA sequence. The obtained results showed complete agreement with some published experimental results on the effects of various mutations [11].

This work can be extended in several directions: (i) Designing similar models for the process of transcription in prokaryotes. (ii) Designing similar models for gene expression in eukaryotes including translation, transcription and splicing. (iii) Applying the developed models to genomes of different organisms. (iv) Using the developed models to obtain new biological findings on gene expression.

3.3. Error correcting codes in the DNA

In contrast to prokaryotes, there seems to be possible redundancy (non-coding DNA) in the genomes of eukaryotes. In humans, for example, protein coding regions comprise only approximately 2% of the whole genome. In fact, it has been observed that the complexity of an organism and its ratio of non-coding to coding DNA is positively correlated [13]. Furthermore, it has been shown that there are several conserved non-coding sequences among species which is a strong indicator of their important functionality [14]. These facts raise the following intriguing questions: (i) Why are higher organisms equipped with so much non-coding DNA (maybe redundancy)? (ii) Can evolution be modelled as an encoder which adds redundancy to the genomic information? (iii) Can one prove the existence of some form of error correcting codes in the structure of the DNA?

DOI: 10.1002/ett

From a coding theory point of view, there is a need to find methods for the detection of a coding structure in a received noisy data sequence whose encoder and decoder are completely unknown. As evolution had a lot of time to optimise its information transmission system, it might be a very complex code. Earlier work has concentrated on gene sequences (i.e. coding DNA) or on prokaryotic organisms with relatively small genomes [2, 3]. However, as the genes are already constrained by the structure of the genetic code, it is unlikely that there are enough degrees of freedom to form an error correcting code. Moreover, simple organisms like prokaryotes have short life cycles and benefit from fast adaptation to a changing environment. That is why they have significant higher mutation rates and smaller genomes than more complex organisms such as eukaryotes. Therefore, it is unlikely that an error correcting code can be found in their genome. As a result of these observations, we believe that if an error correcting code would exist in the DNA, then it is most appropriate to search for it in the conserved non-coding regions of eukaryotic organisms.

In Reference [15], we introduce a novel Kullback–Leibler based method that can identify functional redundancy in the DNA. Evolution is commonly described by a set of parameters $\boldsymbol{\psi}$, representing the phylogeny and a model of mutations [16]. We assume an ancestor input sequence $x[n]$ is transmitted over an evolution channel to output at the receiver the vector $\boldsymbol{y}[n]$. The channel is characterised by the transition probabilities $p_{\mathrm{y}}(\boldsymbol{y}|x;\boldsymbol{\psi})$ conditional on x and parameterised over $\boldsymbol{\psi}$. The channel varies over n as different DNA regions have been subjected to different substitution rates according to the significance of the biological importance of the information they carry. From this point of view, estimating the conservation of a particular DNA region amounts to the estimation of how good the transmission channel was in this region.

From a communication theoretic viewpoint, the maximum conservation is equivalent to the case of noiseless transmission, that is, the transmitted base $x[n]$ is observed unchanged in all components of the receive vector $\boldsymbol{y}[n]$. In this situation, the channel shall be specified by $p_{\mathrm{y}}(\boldsymbol{y}|x;\boldsymbol{\psi}_0)$ and the receive vector $\boldsymbol{y}[n]$ is distributed according to $p_{\mathrm{y}}(\boldsymbol{y};\boldsymbol{\psi}_0)$. For the comparison with the maximum conservation case, we calculate the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$ of the evolutionary model that most likely generated an ensemble of receive vectors $\boldsymbol{Y}[n]$ in a sliding window over the observed data. Then, we calculate the probability mass function $p_{\mathrm{y}}(\boldsymbol{y};\hat{\boldsymbol{\psi}})$ for a column parameterised by $\hat{\boldsymbol{\psi}}$ and compare the estimated distribution with the one corresponding to the maximum conservation process using the Kullback–Leibler distance

$$s[n] = \mathcal{D}\left(p_{\mathrm{y}}(\boldsymbol{y};\hat{\boldsymbol{\psi}}) \,\|\, p_{\mathrm{y}}(\boldsymbol{y};\boldsymbol{\psi}_0)\right) \tag{2}$$

The score $s[n]$ is associated to the column in the middle of the sliding window. Note that a low score corresponds to a good channel and thus a highly conserved region. Figure 3 presents results for the estimation of conservation in addition to the underlying genomic data. Bases are encoded with a unique colour and maximal conserved columns are marked in black. The proposed distance based score signal reflects the different degrees of conservation and, in contrast to earlier methods, does not rely on an accurate model of neutral evolution [14, 17, 18].

This work can be extended in several directions: (i) Expanding the coding theoretic analysis of the conserved non-coding sequences. (ii) Developing methods for the blind detection of error coding structure and applying them to conserved non-coding regions. (iii) Investigating dependencies between coding and non-coding sequences using phylogenetic and information theoretic methods.

3.4. Channel coding structure in the genetic code

The discovery of the mapping of codons to amino acids (known as the genetic code) was a major advance in the field of molecular biology [19]. The genetic code has 64 codons that uniquely map to 20 amino acids which is a redundant mapping. There are many research efforts trying to study the evolution of the genetic code and its optimality properties. The approach used to test optimality is based on generating other mappings of codons to amino acids and trying to compare them with the natural genetic code using physio-chemical metrics such as polarity and hydrophobicity [20].

One can easily show that codons which code for one amino acid are more closely related to one another (in sequence) than they are related to codons that code for other amino acids. In other words, codons that code for one amino acid differ in several cases by just one nucleotide. Thus, single nucleotide mutations (especially in the third location) will often not change the resulting amino acid rather than lead to an error. Investigating protein substitution matrices, another interesting observation is that the smaller the number of codons per amino acid, the higher the self substitution score for that amino acid. A higher self substitution score implies that the amino acid was more often conserved in its location within evolutionary related protein sequences.

Based on the given observations and analysis, the following open research questions can be raised: (i) Can one

DOI: 10.1002/ett

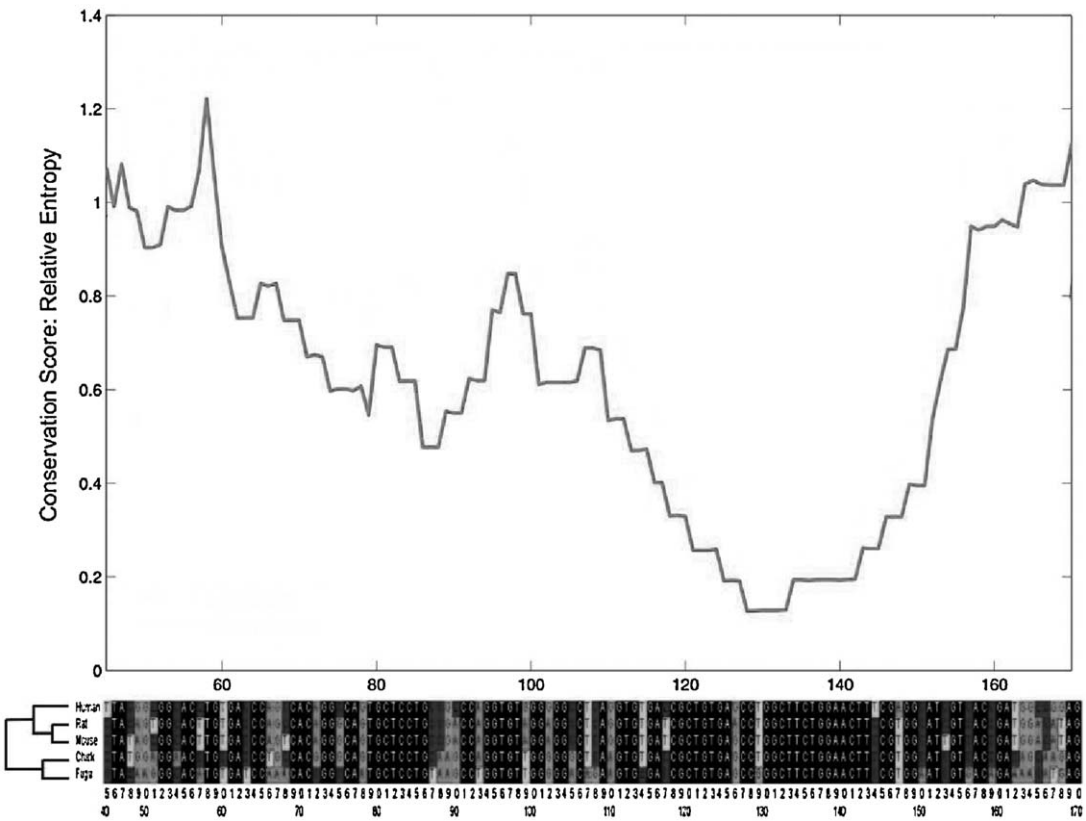

Figure 3. Top, conservation score indicating conserved regions; Bottom, visualisation of the respective genomic data.

justify the structure of the genetic code using channel coding theory? (ii) Can one prove the optimality of the genetic code using channel coding theory? (iii) Is there a relationship between the number of possible codons that result in a given amino acid and the importance of the amino acid? (iv) Is there a relationship between the redundant structure of the genetic code and DNA repair mechanisms?

4. CONCLUSIONS

In this work, we present recent advances in the emerging interdisciplinary field of genomic coding theory. Genomic coding theory deals with applying concepts and techniques from the field of coding theory to problems from the field of molecular biology. The presented applications include source coding for DNA classification and content recognition, channel coding for modelling gene expression processes, existence of error correcting codes in the DNA and channel coding structure in the genetic code.

The following are some practical benefits of this research work: recognition of coding regions in organisms with similar characteristics, gene discovery within a given organism, improving the process of protein synthesis in genetic engineered proteins, etc. As a summary, this work will help in stimulating the interdisciplinary research efforts to apply techniques from the field of communication theory to other problems from the field of genetics.

REFERENCES

1. Yockey H. *Information Theory and Molecular Biology*. Cambridge University Press: Cambridge, 1992.
2. Liebovitch LS, Tao Y, Todorov AT, Levine L. Is there an error correcting code in the base sequence in DNA. *Biophysical Journal* 1996; **71**(3):1539–1544.
3. Rosen GL. Examining coding structure and redundancy in DNA. *IEEE Engineering in Medicine and Biology* 2006; **25**(1):62–68.
4. Battail G. Information theory and error correcting codes in genetics and biological evolution. *Introduction to Biosemiotics*. Springer: New York, USA, 2006.
5. Mac Donnaill DA. Why nature chose A, C, G and U/T: an error-coding perspective of nucleotide alphabet composition. *Origins of Life and Evolution of the Biosphere* 2003; **33**:433–455.
6. May E, Vouk M, Bitzer D, Rosnick D. An error-correcting code framework for genetic sequence analysis. *Journal of the Franklin Institute* 2004; **34**:89–109.

DOI: 10.1002/ett

7. Chen X, Li M, Ma B, Tromp J. DNACompress: fast and effective DNA sequence compression. *Bioinformatics* 2002; **18**(12):1696–1698.
8. Dawy Z, Hagenauer J, Hanus P, Mueller JC. Mutual information based distance measures for classification and content recognition with applications to genetics. In *Proceedings of IEEE ICC 2005*, May 2005.
9. Li M, Badgerand JH, Chen X, Kwong S, Kearney P, Zhang H. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 2001; **17**(2):149–154.
10. Steitz J, Jakes K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in E. coli. *Proceedings of National Academy of Science* 1975; **72**:4734–4738.
11. Dawy Z, Morcos F, Hagenauer J, Mueller JC. Modeling and analysis of gene expression mechanisms: A communication theory approach. In *Proceedings of IEEE ICC 2005*, May 2005.
12. NCBI: National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/.
13. Taft RG, Mattick JS. Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. Eprint q-bio.GN/0401020, January 2004.
14. Siepel A, Bejerano G, Pedersen JS, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Resources* 2005; **15**(8):1034–1050.
15. Hanus P, Dingel J. An alternative method for detecting conserved regions in multiple species. In *Proceedings of the German Conference on Bioinformatics*, October 2005.
16. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology Evolution* 1994; **11**(5):725–736.
17. Margulies EH, Blanchette M, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. *Genome Resources* 2003; **13**(12):2507–2518.
18. Blanchette M. A comparative analysis method for detecting binding sites in coding regions. In *RECOMB'03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, Berlin, Germany, ACM Press, 2003; pp. 57–66.
19. Hayes B. The invention of the genetic code. *American Scientist* 1998; **86**(1):8–14.
20. Freeland SJ, Wu T, Keulmann N. The case for an error minimizing standard genetic code. *Origins of Life and Evolutionn of the Biosphere* 2003; **33**(4–5):457–477.

DOI: 10.1002/ett