

Genome Analysis and Sequences with Random Letter Distribution

Michel Termier

Institut de Génétique et Microbiologie, Université de Paris XI (France)

April 2, 2001

Summary by Mathias Vandenbogaert

Abstract

The information content of genomes of different organisms reflects their mode of physical organisation. For the last decades the wet lab biologist's research interests has been to decipher this information content, with the purpose of extracting useful biological features. The reliability of the information extraction process, mainly based on the textual nature of the underlying messages, was hard to achieve. Therefore, an approach based on the comparison of naturally occurring sequences and randomly generated sequences, is used for discerning the artefacts in sequences and for improving the power of our genome models.

Introduction

The building plan for vegetative life is based on the assembly and catalytic function of proteins and active RNAs. The complete set of instructions that is needed to generate the building blocks of the reproductory system is called a "genome." Any production of living tissue from these building blocks will give rise to an accumulation of secondary metabolites, which are of adverse influence for the survival of the species. The secondary effects of metabolite production are at the basis for the requirement of the genome to be able to respond to the induced environmental changes. To counter this problem, a cell of an organism will only bring to expression those genes that are required at some specific moment in the cell's life cycle. For this purpose, a genome disposes of regulatory systems in the generation processes of building blocks. These systems can be compared to logical gates that are situated in upstream sequences of most information that needs to be processed. This permits a modulation in the usage of information. The genomic information is stocked in a linear fashion, which facilitates the tracking of information. At the time the sequencing of the human genomic sequence is being accomplished, several tasks remain to be addressed:

- the decomposition of the genomic sequence into streams of messages;
- the distinction of these "messages" in contrast to the "non-coding bulk information";
- assignment of biologically significant functions to the messages.

Our bioinformatics team is mainly interested in providing an answer to basically two questions:

1. How can messages be extracted from genomic sequences in order to perform the function assignment task?
2. What is the nature of the message contained within any linear macromolecular structure?

1. First Task: Message Extraction and Function Assignment

The approach consists in observing the known words in the vocabulary of the genome. These known words have been indexed through many years of genetic experiments, with the use of techniques handled in molecular biology wet labs. Through this biology-related knowledge accumulation, the following facts are at the basis for the study of genomic sequences:

- the start and end points (the START and STOP signals) of a nucleic acid sequence correspond to the beginning and to the end of a diffusible product (= protein);
- the information content of a nucleic acid sequence is translated in a unidirectional fashion to the corresponding protein through some basic transcription rules:

$$\text{DNA} \rightarrow \text{messenger (mRNA)} \rightarrow \text{protein};$$

- for the yeast organism, experiments have demonstrated that at least 99 triplets are required between the START and STOP signals, which leads to the coding sequence expression [5]:

$$\text{START}(n^3 \setminus \text{STOP})^{99}(n^3 \setminus \text{STOP})^* \text{STOP},$$

with $n = \{a, c, g, t\}$, $\text{START} = atg$, $\text{STOP} = \{taa, tag, tga\}$;

- by replacing the T-based nucleotides with U, this expression proves to be universally true for the genes describing the intermediate messenger molecules (mRNA) in the steps between DNA and protein;
- for the genomic sequences of higher eucaryotes, the protein-describing sequences are interspersed with non-coding intronic sequences (introns, non-coding bulk information);
- a multitude of other signals exists, regulating the expression of specific coding regions, and responsible for the organism's physiological response in precise environmental conditions.

1.1. Mechanisms for processing signals in messages. There exist mechanisms for processing complex signals, both within eucaryotes as well as within viral species. The *eucaryotic* mechanism is described as *alternative splicing*: a protein-encoding sequence can generate different proteins at the time mRNA is being spliced, according to different translational systems. Sample mechanisms for this group of organisms are read-through (the transcription machinery is reading through and beyond the STOP codon), and hopping (the transcription machinery is skipping the STOP codon and the codons surrounding it). The *retro-viral* mechanism is called *re-encoding*, which implies that different proteins can be obtained at the time the mRNA is being translated. Sample mechanisms for this group are frameshift (the reading frame for translation is changed, which induces an alteration of the encoded amino acids), read-through and hopping. Several features can be conferred to some sequences that are responsible for a frameshift:

1. Slipping sequences (structure X XXY YYZ).
2. A badly positioned classical STOP signal: the ribosome loses his grip on the sequence and gets positioned again in phase -1 .
3. A ribosome-blocking structure.

Regulatory sequences that are responsible for the modulation of DNA transcription in a less error-prone fashion are:

1. *Inhibitor signals*. Their role is to bind proteins so that the RNA polymerase can no longer bind to the sequence to initiate transcription.
2. *Activator signals*. There exists a multitude of signals per protein-encoding sequence, according to the specific function of the protein to be generated.

Usually, these regulatory sequences are short sequences, whose observed frequency is higher (hence unexpected) in comparison to a random word composed of the same letters.

1.2. Modelling a genomic sequence. A Markov model is frequently used for modelling a genomic sequence. The number of sequences that can be generated by this model, increases with the order of the Markov model, and reaches a plateau.

For a Bernoulli-type distribution of the nucleotides, the actual sequence follows a Gaussian distribution. Additionally, when [A+T] increases, the amount of START and STOP signals increases. This implies that the certainty of finding a gene increases.

Regulatory signals are words with biased composition, with respect to the global word distribution of the sequence. These signals have been selected for their properties in the course of evolution. They have been generated according to mechanisms which include random events [2, 3].

1.3. The importance of codon usage biases. In the context of genetic expression, the codon usage bias is correlated with the level of tRNAs available, and with the abundance of protein generated. The level of protein-encoding sequences that are significantly biased is of the order of 20% of the total amount of sequences. Within this respect, several observations have been made:

- the biased structure helps in regulating the transcription turnover [6];
- there is a positional codon bias according to the strand on which the gene is situated [4];
- there is a codon usage bias according to the life cycle of the organism and the cellular location of the metabolic activity [1];
- there is a bias in relation with mRNA stability problems [9];
- some horizontal transfers can have effects on the codon usage [8].

The codon usage bias determining the level of codons corresponding to the amino acids of proteins has a direct effect in the genomic sequence composition of the organism. This bias, which is the result of an interaction of horizontal transfer and metabolic constraints, is at the basis of the selection of efficient proteins. The codon usage bias reveals information about the nucleotide triplet usage of the encoded protein and about the eventual external origin of the sequence in the organism. The significance of the codon usage bias can be evaluated by using weighted linguistics approaches. This consists in heuristically weighting the codons used to encode the amino acids, instead of using an average weight for every amino acid that is encoded by several triplets. This prevents from having resulting frequencies that diverge from the observed values.

Nevertheless, the probability of finding reasonable codon compositions through linguistic methods is fairly low, because:

- global linguistics are calculated on a larger set of oligonucleotides than the number of oligos that determine the proteins;
- the number of codons in a gene equals one third of the number of possible triplets;
- the different genes are built up from codons of different composition, and this is increasing the background noise accordingly.

2. Second Task: Determining the Nature of the Message

Life on any other planet besides Earth can only be detectable for us if it is based on our carbon chemistry. Any sequential organic macromolecule contains constitutional information, if textual organization can be detected within it.

Different approaches exist for the detection of organized information:

1. *Complexity analysis of sequences.* The complexity of sequences is difficult to compute. Ed Trifonov introduced in 1990 the notion of linguistic complexity [7] that reflects the linguistic wealth of a sequence. This complexity is easily computable as $C = \prod_{i=1}^{n-1} u_i$, with u_i the ratio of the words found in a sliding window at position i in a sequence, versus the total number of different words that could possibly be found. Computations are made along

windows, by multiplying the u ratios of words of all possible lengths in the window. This implies that all redundancies are eliminated. The value of C varies from 0 to 1.

2. *Shannon's entropy measure* $H(X) = -\sum_i P(x_i) \cdot \log(P(x_i))$. The entropy $H(X)$ is maximal in the case of a random equiprobable sequence. A reduction in entropy corresponds to a generation of information. This implies that the measurement of the amount of information can be done by:

$$I(X) = H(\text{without message}) - H(\text{with message}).$$

This way, the amount of information can be quantified by comparing a randomly generated Markovian sequence (sequence without message) with a naturally occurring sequence. This measure is related to global information content, but does not give any idea on the distribution of the coding zones of the sequence. It is a common observation in information-bearing texts that coding zones are separated from each other by areas that are more or less deprived of information. If the hypothesis of a non-terrestrial genome makes sense, then its linguistics must respond to the following criterions:

- it must be based on a restricted alphabet;
- it bears coding subsequences that are separated from each other in a way that is recognizable by certain molecules;
- the coding subsequences are likely to share some common characteristics;
- these sequences are constructed using linguistics that can vary from one “genome” to another;
- the reading direction of the sequences is oriented (this should facilitate their regulation);
- the method used to copy the message determines the ordered relation between the coding sequences.

Bibliography

- [1] Chiapello (H.), Ollivier (E.), Landès-Devauchelle (C.), Nitschké (P.), and Risler (J.-L.). – Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Research*, vol. 27, n° 14, 1999, pp. 2848–2851.
- [2] Grantham (R.). – Workings of the genetic code. *Trends in Biochemical Sciences*, n° 5, 1980, pp. 327–331.
- [3] Grantham (R.), Gautier (C.), Gouy (M.), Jacobzone (M.), and Mercier (R.). – Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research*, n° 9, 1981, pp. 43–74.
- [4] Lafay (B.), Lloyd (A.T.), McLean (M.J.), Devine (K.M.), Sharp (P.M.), and Wolfe (K.H.). – Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research*, n° 27, 1999, pp. 1642–1649.
- [5] Oliver (S.G.), van der Aart (Q.J.), Agostoni-Carbone (M.L.), Aigle (M.), Alberghina (L.), Alexandraki (D.), Antoine (G.), Anwar (R.), Ballesta (J.P.), and Benit (P.). – The complete DNA sequence of yeast chromosome III. *Nature*, n° 357, 1992, pp. 38–46.
- [6] Olivier (E.), Delorme (M.O.), and Henaut (A.). – Dos DNA occurs along yeast chromosomes, regardless of functional significance of the sequence. *Comptes rendus de l'Académie des sciences Paris*, n° 318, 1995, pp. 599–608.
- [7] Popov (O.), Segal (D. M.), and Trifonov (E. N.). – Linguistic complexity of protein sequences as compared to texts of human languages. *BioSystems*, n° 38, 1996, pp. 65–74.
- [8] Rocha (E.P.C.), Viari (A.), and Danchin (A.). – Oligonucleotide bias in bacillus subtilis: general trends and taxonomic comparisons. *Journal of Applied Probability*, n° 36, 1998, pp. 179–193.
- [9] Seffens (W.) and Digby (D.). – mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research*, n° 27, 1999, pp. 1578–1584.