UNIVERSITÀ DI PISA
DIPARTIMENTO DI INFORMATICA

TECHNICAL REPORT: TR-05-07

A compression-based approach for coding sequences identification in prokaryotic genomes

Giulia Menconi [a], Roberto Marangoni [b]

RESS: via F. Buonarroti 2, 56127 Pisa, Italy. TEL: +39 050 2212700 FAX: +39 050 2212726

A compression-based approach for coding sequences identification in prokaryotic genomes

Giulia Menconi [a], *Roberto Marangoni* [b]

[a] *Dipartimento di Matematica Applicata, Università di Pisa, Italia*
[b] *Dipartimento di Informatica, Università di Pisa, Italia*

ABSTRACT
Motivation: To identify coding regions in genomic sequences represents the first step toward further analysis of the biological function carried on by the different functional elements in a genome. The present paper presents a novel method for the classification of coding and non-coding regions in prokaryotic genomes, based on a suitable defined compression index of a DNA sequence.
Results: The proposed approach has been applied on some prokaryotic complete genomes, obtaining optimal scores of correctly recognized coding and non-coding regions. Several false-positive and false-negative cases have been investigated in detail, discovering that this approach can fail in the presence of highly-structured coding regions (e.g., genes coding for modular proteins) or quasi-random non-coding regions (regions hosting non-functional fragments of copies of functional genes; regions hosting promoters or other protein-binding sequences, etc.).

Contact: marangon@di.unipi.it (Roberto Marangoni)

1 INTRODUCTION

One of the major problems of the post-genomic era is to ”read” genomic sequences in order to extract all the biological information contained in them. The main task, of course, is to identify the genes, or, in other words, to distinguish which sequences are hosting information codifying for a protein and which are not. Subsequent questions consider possible control relationships between genes, possible multiple splicing procedures (one gene, several products), and all the functional implications related to the localized gene.
Different algorithms have been proposed to detect coding regions, some developed for a specific genome, others of a general applicability.
Apart from the specific differences, the general background hypothesis underlying these approaches is the observation that non-coding regions are often containing repeats, and therefore the information content of non-coding regions is low, if compared with that of coding regions, which tends to that of casual sequences (i.e., coding regions tend to maximize the Shannon entropy per symbol).
A direct consequence of this is that non-coding regions can be easily compressed, where, in general, coding regions cannot. Following this approach, it is necessary to give a precise definition of the compressibility of a given string, and to compute some compressibility index on a genomic sequence, trying to relate the index value to the content of the sequence itself.
The present paper aims at introducing and validating a test for a DNA sequence to be noncoding. This technique is developed in the context of data compression. The data compression schemes to which we refer allow regularities to be associated to compressibility. Therefore, the compression variation within a DNA sequence may be read as information about the regular structures in the sequence. Since in general, non-coding regions show local and global regularities more extensively and intensely than coding regions, this test can be used as a probability measure for a sequence to be non-coding.

2 METHODS FOR FRAGMENT SELECTION

2.1 Fragment Compression

Let $\mathcal{A}^*$ be the space of sequences of finite length whose symbols belong to a finite alphabet $\mathcal{A}$. In the case of genomes, the alphabet is $\mathcal{A} = \{A, C, G, T\}$. If $\sigma \in \mathcal{A}^*$, then its length shall be denoted by $|\sigma|$ and σ^n shall denote the subsequence of σ made of its first n symbols.

Definition 1. (Compression Algorithm) A lossless data compression algorithm is any injective function $Z : \mathcal{A}^* \to \{0, 1\}^*$.

Therefore, a compression algorithm is a reversible coding such that from the original string s may be recovered from the encoded string $Z(s)$. Since the coded string contains all the information that is necessary to reconstruct and describe the structural features of the original string, we can consider the length of the coded string as an approximate measure of the quantity of information that is contained in the original string.

Definition 2. (Computable Information Content) The information content of a finite string $s \in \mathcal{A}^*$ with respect to a compression algorithm Z is defined as

$$I_Z(s) = |Z(s)| \ . \tag{1}$$

The Iinformation Content of a string s is the length[1] of the coded string $Z(s)$.

Moreover, we define another quantity, the complexity of a finite sequence, providing an estimate for the rate of information content contained in it.

Definition 3. (Computable Complexity of a finite string) The complexity of s with respect to Z is the compression ratio

$$C_Z(s) = \frac{I_Z(s)}{|s|} \ . \tag{2}$$

Remark. The experimental results, which will be presented later in this paper, have been obtained by applying a particular compression algorithm, called $CASToRe$. This algorithm is a modification of the Lempel-Ziv compression scheme $LZ78$ (20) and it has been introduced in (3) and its algorithmic properties were studied in (6).

Definition 4. We say that any exon, intron or intergenic region is a functional *fragment* of the genome sequence, following the annotation under GenBank archives (1).

We shall consider the complexity FC of each fragment. The lower it is, the more compressible and regular the fragment is. We have performed a test by selecting any fragment whose complexity is lower than 0.8 out of 1. We denoted this test by **HC**, meaning that any selected fragment is **highly compressible**.

2.2 Sublinearity Index

We shall study the Information Content growth within a fragment. In fact, we have considered how the Information Content varies along each fragment: it is a function of the number of encoded symbols that may be either linear or sublinear. The former means that the sequence is highly irregular, while the latter suggests that there are regular structures within the fragment. This method has been introduced in (14). In the following, σ shall denote any fragment within a genome. The sublinearity index may be defined by means of any adaptive compression algorithm Z, although the experimental results are referred to the algorithm CASToRe.

Let $N = |\sigma|$ be the length of the input sequence σ. Let $\mathcal{P}(\sigma, Z)$ be the parsing of σ with respect to the algorithm Z: $\mathcal{P}(\sigma, Z) = \{\phi_1, \phi_2, \ldots, \phi_t\}$. Therefore, the input string σ is the ordered juxtaposition of phrases ϕ_j's. We use the symbol n_k to indicate the current total number of encoded symbols up to step k of the encoding procedure: $n_k = \Sigma_{j=1}^{k}|\phi_j|$. Due to the fact that $|\phi_k| = n_k - n_{k-1}$, we say that n_k is the parsing index corresponding to the phrase ϕ_k. The Information Content after k steps is then the quantity $I(n_k) = \Sigma_{j=1}^{k} I(\phi_j)$. Obviously, it holds that $n_t = \Sigma_{j=1}^{t}|\phi_j| = N$ and $I(\sigma) = I(N) = \Sigma_{j=1}^{t} I(\phi_j)$. Since the encoding procedure might be not precise in the early steps as well as in the final steps, we fix two bounds defining the restriction of the potential integer value n_j. Let $T_{inf} = 20\%|\sigma|$ be the lower bound and $T_{sup} = 90\%|\sigma|$ be the upper bound. The choice of the bounds will be such that there exist two parsing indexes n_{inf} and n_{sup} such that $T_{inf} \le n_{inf} < n_{sup} \le T_{sup}$.

(a)

(b)

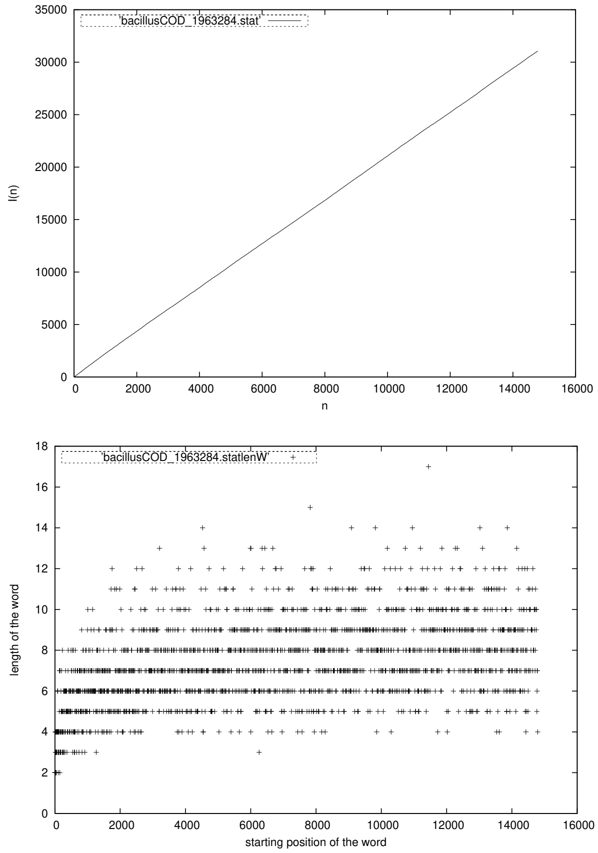

Fig. 1. *B. subtilis* fragment Coding_1963284: (a); (b).

Definition 5. (Sublinearity index of a finite symbol sequence)

Let q_{min}, q_{max} and $q_Z(\sigma)$ be defined as follows:

$$q_{min} = \min_{n_k \in \mathcal{D}} \left\{ \frac{I(n_k)}{n_k} \right\} \ , \ q_{max} = \max_{n_k \in \mathcal{D}} \left\{ \frac{I(n_k)}{n_k} \right\}$$

and

$$q_z(\sigma) = \frac{q_{min}}{q_{max}} \ .$$

[1] Usually, this length is calculated in bit units, but throughout this work we use $\log_4$ instead of $\log_2$ to normalise the numerical results.

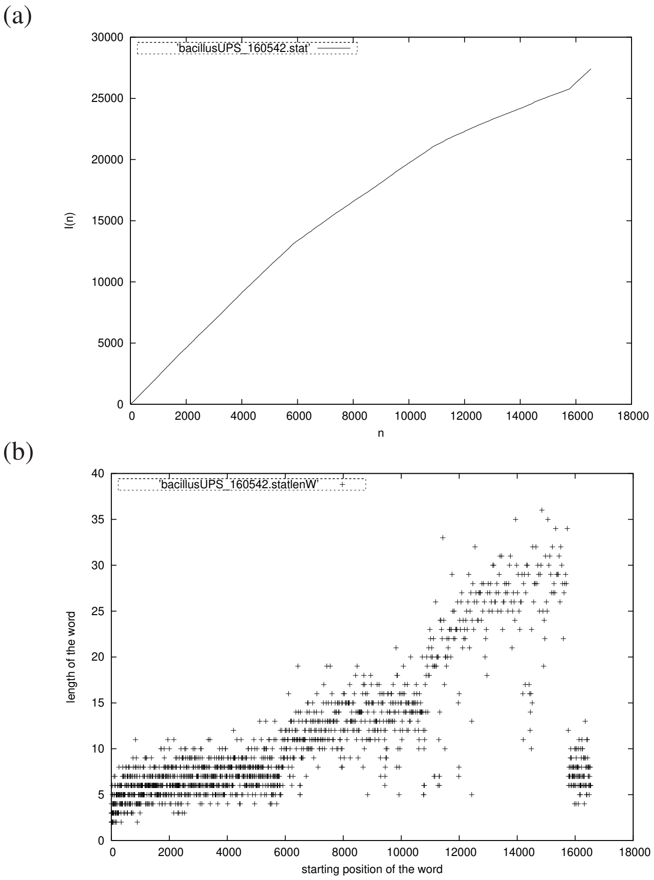

Fig. 2.

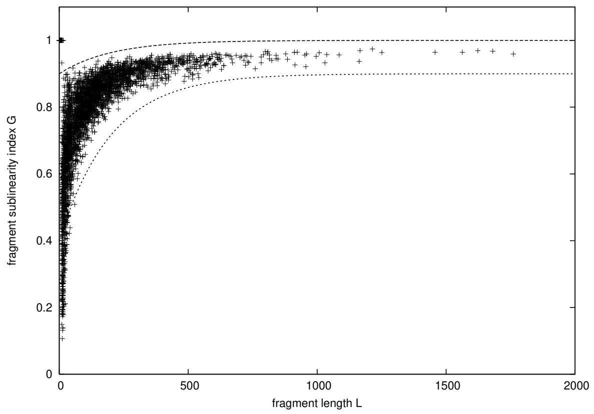

Fig. 3. The relationship between fragment length L and sublinearity index $\mathcal{G}_Z$ in $B.\ subtilis$ (points) for fragments of length up to 2000nt. The upper dashed line is the function $1 - 0.1\exp(-0.005x)$, where the lower dotted line is the function $0.9 - 0.5\exp(-0.005x)$.

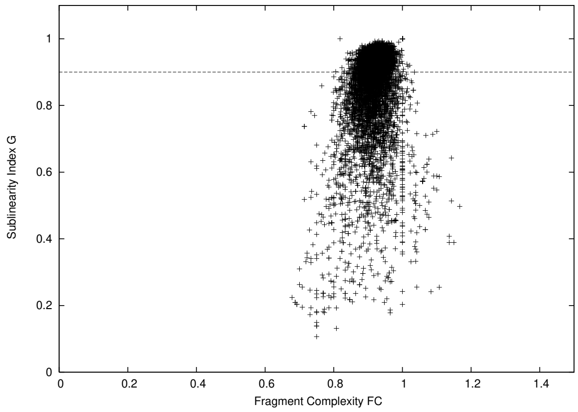

Fig. 4. Comparison between FC and $\mathcal{G}_Z$ in $B.\ subtilis$ fragments. The dashed line is the threshold at $\mathcal{G}_Z = 0.9$.

The sublinearity index $\mathcal{G}_z(\sigma)$ of the input sequence σ with respect to the parsing defined via the algorithm Z is the quantity

$$\mathcal{G}_z(\sigma) = \frac{\log(q_Z(\sigma))}{\log(\frac{n_{sup}}{n_{inf}})} + 1\ . \tag{3}$$

The definition of this index $\mathcal{G}_z$ deserves some comments. Its main characteristic is that it allows a criterion to be established to identify regions where the Information grows sublinearly. The two following main points are definitely true. First, a sublinear growth of Information Content is an indicator of the presence of some regularity in the input sequence and this is much more evident when the index $\mathcal{G}_z$ is significantly smaller than 1. For instance, if $I(n_k) = Cn_k{}^{\gamma}$ with $0 < \gamma \leq 1$ (power-law case), then $\mathcal{G}_z = \gamma$. Second, small values of the index $\mathcal{G}_z$ may correspond to different sublinear information growths − also other than power-law-like − that consequently might be a signal of different underlying dynamics generating the symbol sequences. Preliminar examples are studied in ref. (14).

Performing a sublinearity test, we have extracted any **sublinear fragment (SUB)** whose sublinearity index $\mathcal{G}_z$ is smaller than 0.9. The threshold has been fixed according to the empirical principle that the kind of growth n^{γ} where γ lies in $[0.9, 1]$ is, on a general basis, equivalent to a linear growth, due to the finiteness of the sequences under analysis.

Finally, we have also selected what we called **striking fragment (STRK)**, that is any fragment which is both highly compressible and sublinear.

3 RESULTS

We have applied the SUB, FC, STRK tests described above on some bacterial complete genomes:

- *Methanococcus jannaschii* DSM 2661 (7), version NC_000909.1 GI:15668172
- *Archeoglobus fulgidus* DSM 4304 (11) , version NC_000917.1 GI:11497621
- *Methanobacterium thermoautrophicum* str. deltaH (18), version NC_000916.1 GI:15678031

- *Pyrococcus abyssi* GE5 (13), version NC_000868.1 GI:14518450
- *Aquifex aeolicus* VF5 (9), version NC_000918.1 GI:15282445
- *Escherichia coli* CFT073 (19), version NC_004431.1 GI:26245917
- *Bacillus subtilis* subsp. subtilis str. 168 (12), version NC_000964.2 GI:50812173
- *Haemophylus influenzae* Rd KW20 (10), version NC_000907.1 GI:16271976
- *Mycoplasma genitalium* G-37 (17), version NC_000908.1 GI:12044850
- *Rickettsia prowazekii* str. Madrid E (2), version NC_000963.1 GI:15603881
- *Thermotoga maritima* MSB8 (16), version NC_000853.1 GI:15642775

We have exploited a significance analysis on the tests. We recall that positivity to the test for a region means that it is selected as noncoding. Truth means that the region is noncoding. In particular, we have observed the following indexes.

◇ **Sensitivity Sn** is the ratio $\# \; True \; positive / \# \; true$: how many selected regions are really noncoding.

◇ **Specificity Sp** is the ratio $\# \; True \; negative / \# \; false$: how many not selected regions are really coding.
Notice that Sn and Sp are independent of the fact whether the property of being noncoding is dominant in the analysed collection of regions. Therefore, these two indexes may be used as an unbiased measure to compare tests'results among different genomes.

◇ **Positive Predictive Value PPV** is the ratio $\# \; True \; positive / \# \; positive$: how likely is that a region is noncoding when selected.

◇ **Negative Predictive Value NPV** is the ratio $\# \; True \; negative / \# \; negative$: how likely is that a region is coding when not selected by the test.

We briefly recall the terminology used in the Tables.

A selected and noncoding fragment is true positive.

A selected and coding fragment is false positive.

A not selected and noncoding fragment is false negative.

A not selected and coding fragment is true negative.

Figure 3 shows that the value of the sublinearity index depends on the fragment length and suggests that $\mathcal{G}_Z(L) = O(1 - a \exp(-bL))$ with $a, b \in (0, 1)$.

Figure 4 shows the relationship between the Fragment Complexity FC and the sublinearity index $\mathcal{G}_Z$ for fragments in $B.\ subtilis$ genome. All the studied genomes turn out to share the property that to a sublinear growth may correspond a wide-spreading of FC values. Therefore, the sublinearity index may be more helpful than complexity in identifying distinctive features of regular fragments.

Nr. of fragments		7607			
Nr. noncoding		3814			
Nr. coding		3793			
Test	**Nr.pos.**	**Sn**	**Sp**	**PPV**	**NPV**
SUB	3348	77.79%	89.96%	88.62%	80.11%
true pos.	2967				
false pos.	381				
false neg.	847				
true neg.	3412				
HC	123	3.09%	99.87%	95.93%	50.61%
true pos.	118				
false pos.	5				
false neg.	3696				
true neg.	3788				
STRK	121	3.04%	99.86%	95.87%	50.60%
true pos.	116				
false pos.	5				
false neg.	3698				
true neg.	3788				

Table 1. *Tests'results for B. subtilis*

This is also confirmed by the observation of Tables 1 and 2, which are relative to *B. subtilis* and *E. coli*, respectively. They are taken as examples of the results of the three tests. The general trend is that the SUB test is the most successful and effective. Furthermore, we notice that the case where a fragment is highly compressible but not sublinear is extremely rare (in the Tables there is only one case per genome) and it may be associated either to a true or false negative.

In Table 3, the results for all bacterial genomes concerning the SUB test are shown.

4 CASE STUDIES UNDER SUB TEST IN E. COLI AND B. SUBTILIS

We have considered how sensitivity and specificity of SUB test change as fragment length grows. From Table 4 a huge dependence of test significance on fragment length emerges. In case of short length (up to 200nt and up to 500nt), for both genomes there is a great amount of selected coding regions (false positives), even if the nonocoding regions are selected almost correctly. On the other end, for fragments longer than 1000nt, the results are twofold: for *E. coli* at least the noncoding regions are almost surely predicted, while in *B. subtilis*, though the amount of noncoding regions under this constraint is small, they are extremely irregular, therefore they are definitely excluded by the test. Finally, the test is efficient and reliable in the case of fragments of intermediate length, since

Nr. of fragments	6898				
Nr. noncoding	3450				
Nr. coding	3448				
Test	**Nr.pos.**	**Sn**	**Sp**	**PPV**	**NPV**
SUB	2622	69.68%	93.68%	91.69%	75.54%
true pos.	2404				
false pos.	218				
false neg.	1046				
true neg.	3230				
HC	66	1.88%	99.97%	98.48%	50.45%
true pos.	65				
false pos.	1				
false neg.	3385				
true neg.	3447				
STRK	66	1.88%	99.97%	98.48%	50.45%
true pos.	65				
false pos.	1				
false neg.	3385				
true neg.	3447				

Table 2. *tests'results for E. coli*

Genome	Sn	Sp
Methanococcus jannaschii	76.99%	88.03%
Archeoglobus fulgidus	71.01%	90.17%
Methanobacterium thermoautrophicum	75.93%	92.30%
Pyrococcus abyssi	68.54%	95.83%
Aquifex aeolicus	69.89%	96.68%
Escherichia coli	69.68%	93.68%
Bacillus subtilis	77.79%	89.96%
Haemophylus influenzae	77.03%	92.95%
Mycoplasma genitalium	53.66%	93.95%
Rickettsia prowazekii	62.92%	91.39%
Thermotoga maritima	67.90%	92.01%

Table 3. *Comparison of SUB test's significance for several bacterial genomes*

when the regions are up to 1000nt long, the significance of the results is high.

In order to better characterise the features of the fragments that were wrongly selected under SUB test, first we selected some deeply regular coding fragments from *B. subtilis* and *E. coli* genomes. Second, we extracted some noncoding irregular fragments.

By means of homology searches in databanks, we studied the structure of these sequences and try to infer their biological role.

E. coli		
Constraint	**Sn**	**Sp**
fragments up to 200nt long	83.27%	12.86%
fragments up to 500nt long	72.93%	70.35%
fragments up to 1000nt long	70.05%	87.97%
fragments longer than 1000nt	20.00%	99.88%
B. subtilis		
Constraint	**Sn**	**Sp**
fragments up to 200nt long	89.50%	17.12%
fragments up to 500nt long	80.57%	64.89%
fragments up to 1000nt long	78.14%	83.69%
fragments longer than 1000nt	4.76%	99.86%

Table 4. *Significance analysis for E.coli and B. subtilis under SUB test with different constraints on fragment length.*

4.1 False positives: coding regular fragments

Coding sequences falsely marked as non-coding codify for very particular kinds of proteins: modular proteins or proteins consisting of only a few species of aminoacids.

These proteins are, of course, not common, and significantly deviating from the "average" protein, but they are actively expressed and some of them are highly preserved (even in higher organisms). There are no doubts, therefore, that these sequences are "coding", but the very particular structure of them, which consists in repeats of a short aminoacid motif, can get false results from our test.

4.2 False negatives: noncoding irregular fragments

The fragments erroneously-classified as coding by SUB test mainly belong to two classes:

1. Pseudo- or non-functional-genes or fragment of a gene, which probably derive from a gene duplication event, in which the copy has failed to produce a complete and functional gene.

2. Non-coding sequences which are strictly related to transcription regulation: enhancers, promoters or promoter fragments, which have lost their respective genes.

In the first case, our test indicates a coding region since the test is not able to perform a functional evaluation and therefore it cannot distinguish between an actual and active gene and "an evolutive ghost".

In the second case, the sequences are non-coding, because they don't express a gene, but their structure is quasi-random, as we can observe in coding regions. Therefore our test can detect them as coding. Usually, these sequences, when functionally active, are physically located close to the controlled gene, and therefore the classification error is very low.

To sum up, the performance of our proposed method for coding regions classification is very good, and most of the cases in which it fails can be explained in terms of a functional role played by the examined sequence.

This represents a strong suggestion for further improvement of the present work, which can be obtained by add methods for functional evaluation (e.g., search for promoter regions in the upstream of a predicted coding sequence) to the coding/non-coding classification.

REFERENCES

[1]http://www.ncbi.nlm.nih.gov/Genbank/

[2]Andersson,S.G. et al., "The genome sequence of Rickettsia prowazekii and the origin of mitochondria", *Nature* 396 (6707), 133-140 (1998)

[3]Argenti F., Benci V., Cerrai P., Cordelli A., Galatolo S., Menconi G., "Information and dynamical systems: a concrete measurement on sporadic dynamics", *Chaos, Solitons and Fractals*, **13**, 3, 461–469 (2002).

[4]"The Arabidopsis Genome Initiative", 2000, *Nature*, 408:796-815

[5]Bell T., Witten I. H., Cleary J. G., *Modeling for text compression*, ACM Computing Surveys, **21**, 557–591 (1989).

[6]Bonanno C., Menconi G., "Computational information for the logistic map at the chaos threshold", *Disc. Cont. Dyn. Syst.- B*, **2**, no.3, 415–431 (2002).

[7]Bult,C.J. et al., "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii", *Science*,**273** (5278), 1058-1073 (1996)

[8]Cover T. M., Thomas J. A., *Elements of Information Theory*, Wiley (1991).

[9]Deckert,G. et al., "The complete genome of the hyperthermophilic bacterium Aquifex aeolicus", *Nature* 392 (6674), 353-358 (1998)

[10]Fleischmann,R.D. et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd", *Science* 269 (5223), 496-512 (1995)

[11]Klenk,H.P., et al., "The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus",*Nature*,**390** (6658), 364-370

[12]Kunst,F. et al., "The complete genome sequence of the gram-positive bacterium Bacillus subtilis", *Nature* 390 (6657), 249-256 (1997)

[13]Lecompte,O., et al., "Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea", *Genome Res.* 11 (6), 981-993 (2001)

[14]Menconi G., "Sublinear growth of Information in DNA sequences", *Bull. Math. Biol.*, in press.

[15]Myers G., "Whole-genome DNA sequencing", *" Computing in Science & Engineering*, **1**, 3:33–43 (1999).

[16]Nelson,K.E. et al.,"Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima", *Nature* 399 (6734), 323-329 (1999)

[17]Peterson,S.N. et al., "Characterization of repetitive DNA in the Mycoplasma genitalium genome: possible role in the generation of antigenic variation", *Proc. Natl. Acad. Sci. U.S.A.* 92 (25), 11829-11833 (1995)

[18]Smith,D.R., et al., "Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics",*J. Bacteriol.*, 179 (22), 7135-7155 (1997)

[19]Welch,R.A. et al., " Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli", *Proc. Natl. Acad. Sci. U.S.A.* 99 (26), 17020-17024 (2002)

[20]Ziv J., Lempel A., "Compression of Individual Sequences Via Variable-Rate Coding", *IEEE Transactions on Information Theory*, **24**, 530–536 (1978).