ELSEVIER

Gene 300 (2002) 97–104

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

A simple and species-independent coding measure

P. Carpena[a,*], P. Bernaola-Galván[a], R. Román-Roldán[b], J.L. Oliver[c]

[a]*Departamento de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, 29071 Malaga, Spain*
[b]*Departamento de Física Aplicada, Universidad de Granada, Granada, Spain*
[c]*Departamento de Genética and Inst. de Biotecnología, Universidad de Granada, Granada, Spain*

Received 21 December 2001; received in revised form 6 July 2002; accepted 18 September 2002

Abstract

We present a coding measure which is based on the statistical properties of the stop codons, and that is able to estimate accurately the variation of coding content along an anonymous sequence. As the stop codons play the same role in all the genomes (with very few exceptions) the measure turns out to be species-independent. We show results both for prokaryotic and for eukaryotic genomes, indicating, first, the accuracy of the measure, and, second, that better prediction is achieved if the measure is applied on homogeneous, isochore-like sequences than if it is applied following the standard moving window approach. Finally, we discuss on some of the possible applications of the measure.

Keywords: Coding measure; Genomes; Species-independent

1. Introduction

The great number of professional gene-finding programs (GFP) currently available (http://igs-server.cnrs-mrs.fr/igs/banbury/programs.html) are key tools in annotating the current deluge of genome sequence data. Most of the algorithms on which these software tools are built are accurate enough for bacterial genomes but they all show serious flaws when applied to eukaryotes, thus making the computational gene detection still an open problem (Burset and Guigó, 1996; Guigó, 1997; Guigó et al., 2000). This may be due in part to the intricate complexity of eukaryotic gene structure, hard to replicate by simple gene models ignoring alternative splicing, but also to the fact that many assumptions in these, sometimes sophisticated, algorithms are not met by eukaryotic DNA sequences.

Two major drawbacks shared by all GFP are that: (i) they use some parameters that need to be fitted, and therefore previous training on any well-known sequence subset is needed; and (ii) these parameters are commonly species dependent, and thus they need to be determined separately for each different organism.

In this paper, we present a method able to determine in a simple way the relative density of coding regions within anonymous DNA sequences. It is based on the statistical properties of the distribution of distances between in-frame stop codons (Senapathy, 1986), known to vary between genome regions differing in both guanine-cytosine (GC) content and gene density (Merino et al., 1994; Boldögkoi et al., 1996; Oliver and Marín, 1996). The method presents some interesting properties: (i) it is completely species independent; (ii) it is non-parametric, and therefore it does not need any previous training; and (iii) it is very simple from the computational standpoint, and thus it can be easily applied to get some insight about the possible coding content in an anonymous sequence. This type of information can be very valuable when facing large anonymous sequences prior to the use of professional gene finders.

2. Data and methods

Here, we present results for different sequences. In particular, we have considered the complete genome of *Bacillus subtilis*, the scaffold AE002690 of *Drosophila melanogaster* (all retrieved from GenBank) and the largest contigs available for every *Homo sapiens* chromosome

Abbreviations: bp, base; kb, kilobase; Mb, megabase; G + C or GC: molar fraction of guanine and cytosine in DNA.

* Corresponding author.
E-mail address: pcarpena@ctima.uma.es (P. Carpena).

0141-933/02/$ - see front matter
PII: S0378-1119(02)01041-7

(retrieved from the NCBI: ftp://ncbi.nlm.nih.gov/genomes/H_sapiens).

2.1. The coding measure

The method we propose here is based on the statistical properties of the stop codons. It is well known that two (possible) stop codons are never in the same reading frame within a coding region, because otherwise the coding region will be prematurely truncated. In other words, a coding region, either a monoexonic gene or an exon, must be placed within an open reading frame. This restriction for the stop codons should be reflected in the statistical properties of the distribution of distances between consecutive and in-frame stop codons, and therefore coding-rich sequences and coding-poor sequences should present clear differences in such distribution. The reason is clear: for non-coding DNA sequences, in principle there is no restriction for the position of the stop codons, or, in other words, they can appear at random. In contrast, for coding DNA sequences, the positions of the stop codons are restricted because of the presence of coding regions, and for a particular coding region, no stop codons can be found in its reading frame (although they can be found in any of the remaining independent frames).

These facts suggest a simple model accounting for the properties of non-coding DNA. As the stop codons can appear at random, it is straightforward to determine the distribution of distances between consecutive in-frame stop codons in non-coding DNA.

In general, let us define the probability of finding a stop codon within a certain sequence, p_{stop} as

$$p_{\text{stop}} = p_{\text{TAA}} + p_{\text{TAG}} + p_{\text{TGA}} \tag{1}$$

i.e. the sum of the probabilities of the three different stop codons. These individual probabilities can be determined simply by counting the number of stop codons appearing in the three independent reading frames of the DNA strand. Once p_{stop} is known, if the stop codons are distributed at random, then the probability of finding two consecutive in-frame stop codons at distance d is given by the geometric distribution, i.e.

$$p_{random}(d) = p_{\text{stop}}(1 - p_{\text{stop}})^{(d-1)} \tag{2}$$

where the distance d is measured in codons. Note that Eq. (2) represents the expected distribution in a non-coding DNA sequence.

Thus, given a particular and unknown DNA sequence, we can calculate the real distribution of distances between consecutive in-frame stop codons, $p_{\text{real}}(d)$, and compare this distribution with the expected one for a purely non-coding sequence, $p_{\text{random}}(d)$. The latter distribution is calculated, according to Eq. (2), once p_{stop} is determined from the real sequence. In this way, both distributions have the same probability for the stop codons, but p_{real} reflects the spatial distribution in the real sequence, and p_{random} the expected distribution in a purely random, non-coding sequence. We propose that the greater the difference between the distributions $p_{\text{real}}(d)$ and $p_{\text{random}}(d)$, the higher the coding concentration in the unknown sequence.

To quantify the difference between the two distributions, we consider for simplicity the well-known χ^2 distance, and we expect the coding content in an unknown DNA sequence to be related to the distance between $p_{\text{real}}(d)$ and $p_{\text{random}}(d)$. Actually, we expect an increasing monotonic dependence between the χ^2 distance and the coding content of the sequence. This expectation is confirmed in the results obtained (see below), from which it can be inferred a linear dependence between χ^2 and coding content, at least in a moderately large range of GC values.

Therefore, we define the coding measure used in the rest of the paper simply as χ^2 (p_{real}, p_{random}). Note that we choose χ^2 distance just because its simplicity and also because it is probably the most well-known distance between probability distributions. Nevertheless, we would like to note that χ^2 is not the only possible choice and many others are possible (Jensen–Shannon divergence, Kolmogorov–Smirnov test, etc.).

2.2. Approaches to use the coding measure

A first approach to use our coding measure consists of studying the coding-content variations along a given sequence. A possible way to do it is to use a moving-window plot: we choose a window of certain size, and calculate the coding measure in the sequence contained within the window. We assign the result to the central nucleotide of the window, move the window a certain number of nucleotides (step) and repeat the procedure. The resulting plot represents the fluctuations of our measure as a function of the position in the sequence, and we expect these fluctuations to correspond also to the coding content fluctuations in the sequence. Actually, control experiments can be performed: if the measure is tested in an annotated sequence, the moving-window plots of the measure can be compared with the window plots of the real coding content extracted from the annotations to check the validity of our measure. This real coding content is calculated simply by adding up the sizes of the exons and monoexonic genes contained in the window, and dividing the total sum by the window size. In Section 3 we show this type of control experiment.

Another possibility to measure coding-content variations within a certain sequence is to use homogeneous regions instead of moving windows. Eukaryotic genomes are known to be complex systems made up of fairly homogeneous segments of different composition called isochores (Macaya et al., 1976; Bernardi et al., 1985; Bernardi, 2000). In a recent work (Oliver et al., 2001), we developed a computational method able to detect long homogeneous genome regions (isochore-like) at the sequence level. Once

this algorithm is applied, the sequence is completely partitioned into homogeneous and non-overlapping segments. Thus, we can use these homogeneous segments instead of moving windows to calculate our measure. However, since isochores show a very wide distribution of sizes (Oliver et al., 2001, 2002), and, since our statistical measure works only when applied to large enough sequences (the reason being that a large sequence is needed for a reliable estimation of the distribution $p_{\mathrm{real}}(d)$), we will use a certain size threshold before computing our measure on isochores. It is convenient to point out here that this threshold in the sequence size used in the isochore approach is also needed when moving windows plots are considered, the reason being the same: the obtainment of a reliable distribution $p_{\mathrm{real}}(d)$, without excessive numerical fluctuations due to low sampling. We have tried several window (segment) sizes, and we have found empirically that an appropriate size threshold is about 500 kb. Window (segment) sizes larger than this threshold do not produce significant differences, so in the result sections we use sizes of 500, 600 kb and 1 Mb just for convenience (i.e. for not very large sequences, like *B. subtilis*, we use 500 kb and we use larger sizes for larger sequences).

Finally, the measure can be used also to compare the coding content of different sequences, and not only the fluctuations in coding content within a sequence. To proceed, one can calculate the measure in a whole sequence (thus obtaining a single number) and compare this number with the results for other sequences. In this way, a ranking of coding richness can be established.

3. Results

To test the validity of our coding measure, we have selected DNA sequences corresponding to different organisms: a bacterial genome (*B. subtilis*) and several examples of eukaryotic sequences (extracted from the genomes of *D. melanogaster* and *H. sapiens*).

3.1. B. subtilis

In general, the prediction of coding content in bacterial genomes is not a difficult task, because the abundance of genes is high, and also because introns are absent from bacterial genomes. Therefore, bacterial genomes are an excellent test for our measure, because they are fairly well annotated and thus the comparison between real coding content and our results is direct. Therefore, we have selected *B. subtilis* as an example, but the performance of our measure is quite similar in other bacterial genomes.

In Fig. 1 we present the real coding content in both the Watson and the Crick strands of *B. subtilis* and our prediction in the same sequences. These results have been obtained using a moving-window approach. In particular, for the plots in Fig. 1, we use 500 kb of window size and 20 kb of step. We observe that *B. subtilis* shows a striking fluctuation in coding content between the beginning and the end of the sequence in both strands, revealing the strand asymmetry.

Fig. 1 shows an excellent correlation between the real coding content and the prediction with our measure, and thus these results validate the assumptions of the model on which our measure is based. In addition we have performed a control test based on the GC content in the moving windows approach (see inset of Fig. 1 (bottom plot)) showing that little or no correlation can be found between the GC content and the coding density, and then that our measure is more convenient to detect coding richness in this case.

3.2. D. melanogaster

We have also tested our measure in the scaffold AE002690 (16,346,801 bp) of *D. melanogaster*. Fig. 2 shows the results in this sequence in a moving window (window size = 1 Mb, step size = 50 kb) for both the known coding content read from the annotations and our prediction. The correlation between the two results is fairly

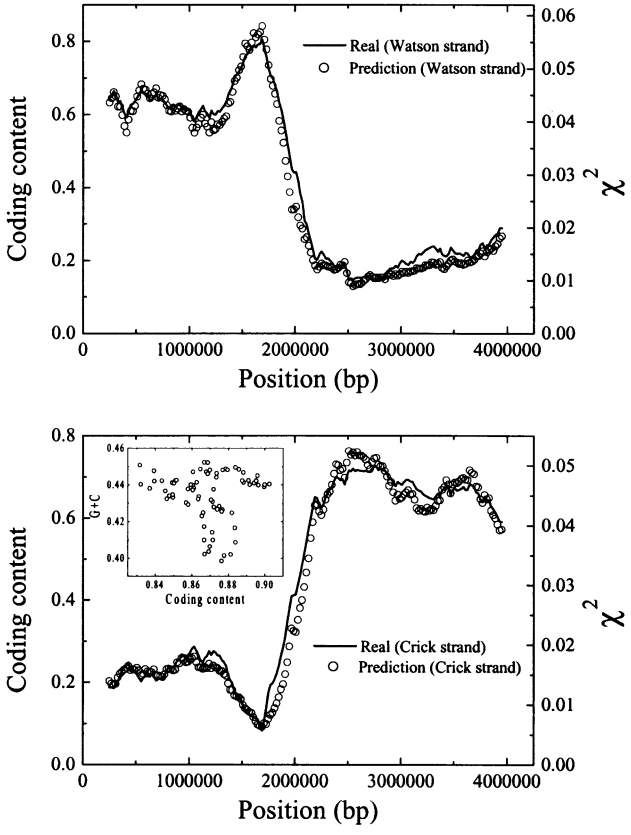

Fig. 1. Coding densities calculated from the annotations (left vertical axis) and our prediction (right vertical axis) in the Watson and Crick strands of *B. subtilis*. All of them have been obtained with a moving window of 500 kb and a step of 20 kb. Inset: Control experiment in which it is represented the GC content versus the coding content of each window for the *B. subtilis* genome.

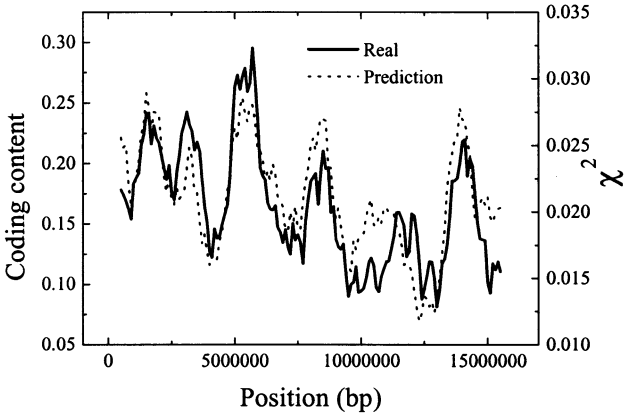

Fig. 2. Coding content obtained from the annotations (solid line) and our prediction (dotted line) for the scaffold AE002690 of *D. melanogaster*. The left vertical axis corresponds to the coding content, while the right vertical axis corresponds to our measure. Parameters used: window size = 1 Mb; step size = 50 kb.

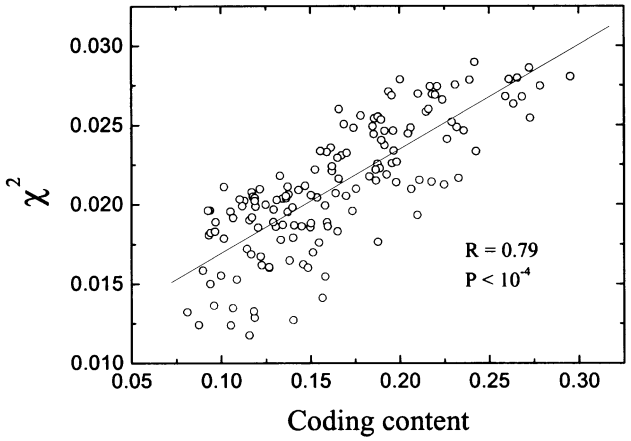

Fig. 3. Coding measure versus the coding content from the annotations (see Fig. 2) for the *D. melanogaster* scaffold AE002690 (16,346,801 bp). 0 5,000,000 10,000,000 15,000,000 20,000,000.

good as can be seen by visual inspection, especially in the regions with higher and lower coding content.

To quantify the degree of correlation between the two results, in Fig. 3 we represent for each window the known coding content as a function of the prediction, this giving a good linear fit ($R = 0.8$, $P < 10^{-4}$). Note that in this case genes may be composed of several exons, in contrast to the case of *B. subtilis*, but our measure nevertheless gives accurate results. However, note that the real coding content is obtained mainly from predictions carried out by professional gene finders, and thus our measure and the 'real' coding content may not be completely independent (see also Section 4).

3.3. H. sapiens: moving-window approach

Due to the recent availability of large and annotated human contigs, it is possible to perform a comparative study between the annotations and the results given by our method. We choose the human chromosome 22 largest contig (NT_011520) and the human chromosome 21 largest contig (NT_011512) as representative examples, and as in previous cases, we adopt a moving-window approach to plot both the known coding content and our prediction as a function of the position in the sequence (Fig. 4, top and bottom plots, respectively). The plots show great coding content fluctuations in the sequences, but again we find a positive correlation between the annotations and our prediction.

These correlations can be properly quantified (as we did before) if we represent the coding content in each window versus the prediction in the same window. We show in Fig. 5 (top plot) the result for the contig of human chromosome 22, where a good correlation coefficient ($R = 0.77$, $P < 10^{-4}$) results. A control test based on the GC content has also been carried out in the same sequence, following the same approach: we represent the coding content in each window versus the GC content in the same window (Fig. 5, bottom plot). Although a good correlation between both magnitudes is obtained ($R = 0.68$, $P < 10^{-4}$), our measure outperforms this result.

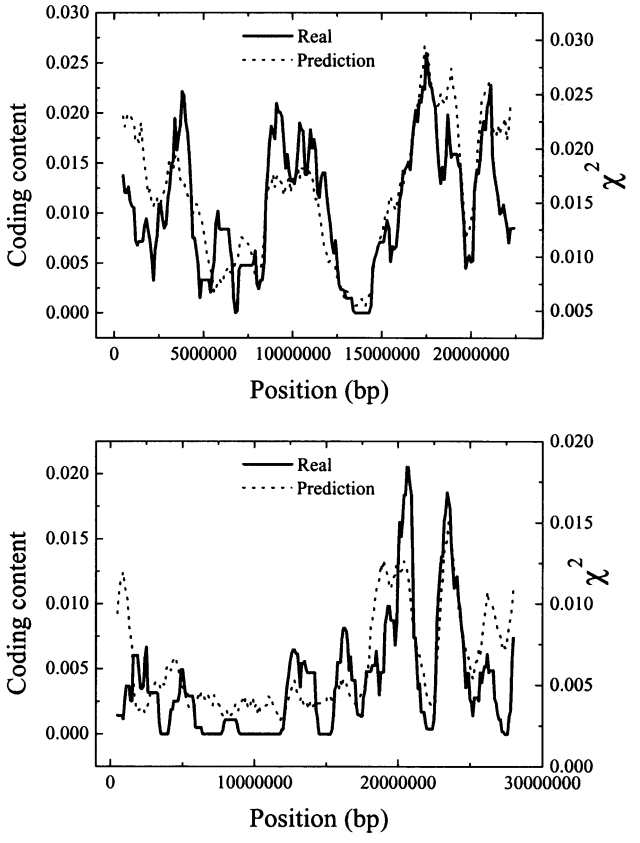

Fig. 4. Coding content from the annotations (left vertical axis) and our prediction (right vertical axis) for the longest contig (NT 011520) of *H. sapiens* chromosome 22 (top plot) and for the longest contig (NT 011512) of *H. sapiens* chromosome 21. Parameters used: window size = 1 Mb; step size = 50 kb.

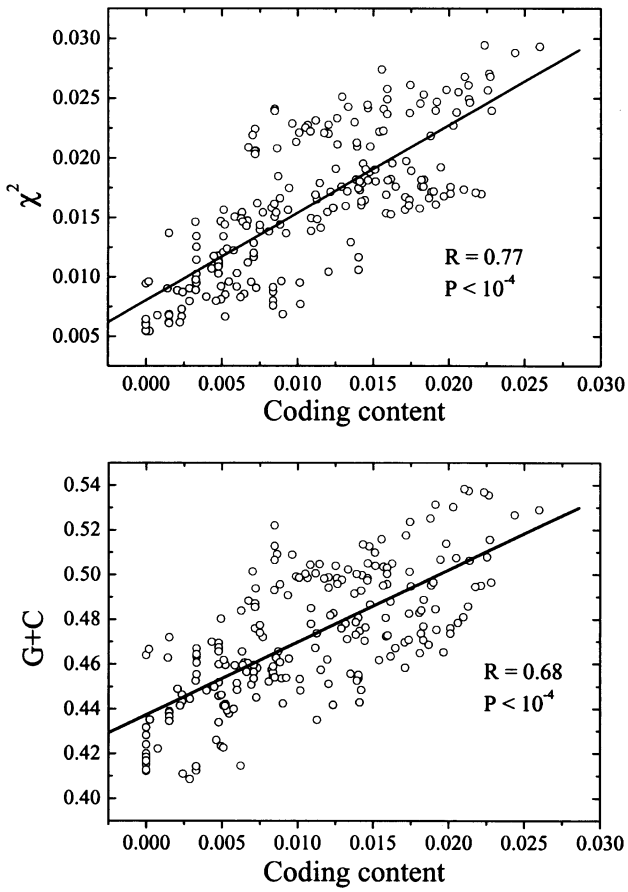

Fig. 5. Top plot: Coding measure versus the coding content from the annotations (see Fig. 4) for the largest contig (NT 011520) of *H. sapiens* chromosome 22. Bottom plot: G + C level versus the coding content for the same sequence. Both plots have been obtained with a moving window of 1 Mb of size, and a step size of 50 kb.

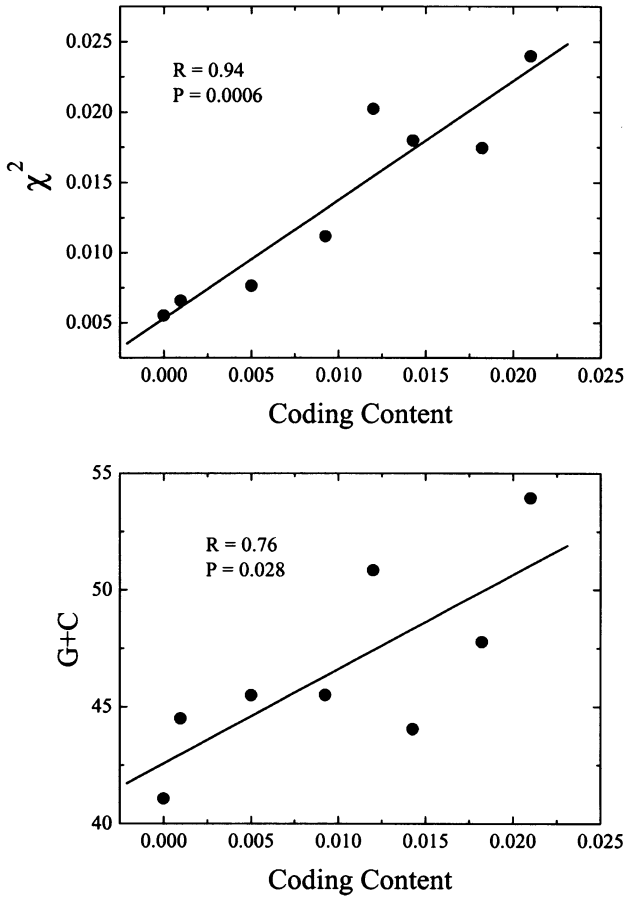

Fig. 6. Top plot: Coding measure versus the coding content obtained from the annotations corresponding to the isochores greater than 600 kb of the largest contig (NT 011520) of *H. sapiens* chromosome 22. Bottom plot: Isochore GC content versus real coding content in the same isochores as in the top plot. The eight isochores selected for these plots have an average size of 1.06 Mb, and represent a 35% of the total size of the contig.

3.4. H. sapiens: isochore approach

Our coding measure can also be calculated in large homogeneous segments (isochore-like), which we delimit at the sequence level by using our algorithm (Oliver et al., 2001). As we commented above, in this approach a minimum isochore size is required to ensure the statistical reliability of the results. In Fig. 6 (top plot), we take a size threshold of 600 kb and plot the coding content (from the annotations) and our estimate for the isochores of *H. sapiens* chromosome 22 largest contig (NT_011520) (see ref. Oliver et al., 2001, for further details on the position and size of the isochores considered). The resulting eight isochores larger than 600 b present an average size of 1.06 Mb, and represent the 35% of the total size of the contig.

Note that in this case, the correlation between real coding content and prediction is clearly higher ($R = 0.94$) than in previous cases, obtained through moving windows plots. Note that the results obtained in Fig. 6 (top plot) are directly comparable to the results shown in Fig. 5, because the window size used in the latter (1 Mb) is almost identical to the average isochore size in the former (1.06 Mb). We discuss the reason for this improvement in the last section of this paper. We have performed again a control experiment based on the relation between the GC content of the isochore, and the coding content of the isochore. In Fig. 6

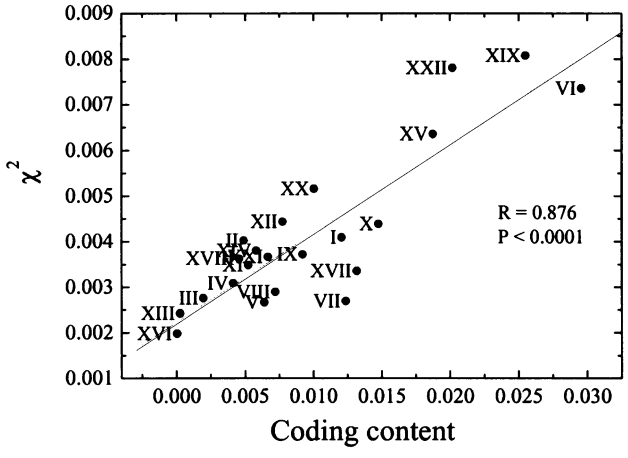

Fig. 7. Coding measure versus the coding content obtained from the annotations for the largest contigs available of the *H. sapiens* chromosomes.

(bottom plot) we represent the correlation between both magnitudes. Although a positive correlation is found ($R = 0.77$), our measure clearly outperforms this control.

3.5. H. sapiens: comparison between different sequences

Up to now, either with the moving-window approach or with the isochore approach, we have studied the results of our measure in determining the coding-content fluctuations within a certain sequence. However, as we suggested before, a test can be performed to compare the coding richness in a set of sequences. As an example, in Fig. 7 we compare the coding content from the annotations and our prediction for the set of sequences formed by the largest contig of each of the human chromosomes. Despite the great differences among all the contigs shown in Fig. 7 (in size, in composition, in internal heterogeneity, …) the measure properly quantifies their relative coding richness.

4. Discussion

We have presented a simple coding measure based only on the statistical properties of stop codons – the behavior of the distribution of distances between consecutive and in-frame stop codons. As the stop codons play the same role in all the genomes with the standard genetic code, and in-frame stop codons cannot be found within a coding region, the coding measure we have developed turns out to be species-independent and does not require prior training. Nevertheless, note that the size of an open reading frame is a common filter used by professional gene-finding programs (which, in addition, do many other tests to decide whether a certain sequence is a coding region or not). The difference with respect to our measure is that these programs decide on an individual basis whether an open reading frame is acceptable or not; they fix a certain size threshold (determined from previous training) and accept (reject) open reading frames greater (shorter) than the threshold. In our case, we obtain the size distribution of open reading frames, and compare it to the expected distribution in a random sequence with the same probability for the stop codons. This latter case presents three advantages: (1) the comparison between probability distributions is always more precise from the numerical point of view than the comparison of the probability for only one event. (2) We do not fix any threshold, and therefore no training is required. (3) The probabilities of the stop codons are estimated directly from the sequence considered. Hence, this probability changes automatically as a function of the sequence composition, implying that the random sequence used to model non-coding DNA, and to compare with the real sequence, also changes automatically. This fact is crucial; for example, in GC-rich sequences, large open reading frames can exist by chance, (simply because the stop codons are GC-poor; see Oliver and Marín, 1996) and not because these open reading frames correspond to coding regions. Our measure filters out these large but randomly generated open reading while taking into account the local probability of the stop codons.

Nevertheless, note that we do not claim that the measure presented in this paper outperforms the results of professional gene-finders: on the one hand, it is not true, and on the other hand, this is not our intention. Actually, these programs obtain extremely accurate results in the prediction of individual genes. Our aim is to develop a simple measure to obtain with easiness the coding content variations along anonymous sequences, which can be a valuable first test prior to the use of professional gene finders.

We have tested our measure both in bacterial and in eukaryotic genomes following different approaches. In general, we find a rather good correlation between our prediction and the coding content determined from the annotations of the sequences considered. In principle, these results are expected in bacterial genomes, where gene structure is very simple, and we find that the correlation between real coding content and our prediction is excellent, as in the *B. subtilis* case shown above. However, we have also found quite good results when considering eukaryotic sequences, despite the inherent complexity of the gene structure in this latter case. Note, however, that this good agreement between our results and the 'real' coding content may be partly due to the fact that the real coding content is obtained in the majority of cases from predictions of professional gene-finders, and thus our results and the coding content may not be completely independent a priori.

We have compared our prediction and the annotations in moving-window plots and isochore plots when studying the fluctuations of coding content along a certain sequence. As a general result, our prediction in the isochore approach is better than in the moving-window approach. The main reason for this improvement is homogeneity; that is, our model for non-coding DNA is based on the hypothesis that non-coding DNA can be modeled by a random distribution of stop codons, which appear in the sequence with a certain probability p_{stop}. This probability p_{stop} is assumed to be constant in the sequence considered. However, a real sequence is in general far from homogeneous – fluctuations in local GC content lead to different local values of p_{stop}. When using a moving-window approach, the window can contain a highly heterogeneous fragment of the total sequence, and therefore different local values of p_{stop}. Nevertheless, the calculation averages the different values of p_{stop} in the corresponding window to assign a single p_{stop}, this fact being a source of prediction errors. By contrast, when we consider the sequence divided into non-overlapping homogeneous segments (isochore-like), the method used to obtain these segments ensures their homogeneity. This homogeneity is granted up to a certain degree of statistical confidence, because fluctuations in composition also exist within isochores (Clay and Bernardi, 2001), which may be due in part to the existence of correlations

(Bernaola-Galván et al., 2002). This relative homogeneity makes p_{stop} to be approximately constant within each isochore-like segment, which finally leads to a better prediction than in the moving-window case.

4.1. Possible applications of the coding measure

The properties of our coding measure make several direct applications possible. Here, we cite three.

(1) The measure can be used as a preliminary test prior to the use of a professional gene-finding program to annotate a sequence. Our measure is specially suitable to detect the coding-content fluctuation within a sequence, and thus to find the regions of the sequence rich in coding DNA in a simple and fast way even in very large sequences. The data provided by our measure can be very useful as preliminary information prior to the systematic gene prediction task performed by professional gene finders.

(2) Recent studies (Devos and Valencia, 2001; Skovgaard et al., 2001) suggest that microbial genomes are often wrongly annotated, and that, in general, the number of genes in such genomes are systematically overestimated. The reason is that often, almost all open reading frames were considered genes, regardless of their lengths. Nevertheless, as discussed above, depending on the composition, large open reading frames can appear simply by chance, without corresponding to real genes. This gene overestimation can reach values as high as 10% in some cases. However, our measure takes into account the probability of having large, open randomly-generated reading frames, because the probability p_{stop} changes automatically with the composition. Thus, by careful comparisons between our results and the coding content extracted from the annotations, regions with overestimations of genes can be detected.

(3) The debate concerning the total number of human genes continues. From initial estimations of about 30,000–40,000 genes (Lander et al., 2001; Venter et al., 2001), recent studies conclude that a more accurate value may be twice the previous estimate: between 65,000 and 75,000 genes (Hollom, 2001). Our measure could also be used to give at least a rough estimation of this number if the proportionality factor between our measure and the real coding content is accurately determined. Work on such an estimation is under way (Carpena et al., 2002).

5. Conclusions

We have presented a simple measure able to predict with confidence the coding content fluctuations within a sequence for both bacterial and eukaryotic genomes. In addition, the measure can also be applied to compare the coding content of different sequences. The measure is species independent, because is based on the statistical properties of the stop codons, which have the same function in all the organisms with the standard genetic code. We have shown that, although moving-window plots are appropriate to use the measure, a better prediction is achieved if the measure is tried in homogeneous, isochore-like sequences. Finally, we have discussed some of the possible applications of the coding measure, such as preliminary tests prior to the use of professional gene finding programs, a possible tool to correct the gene overestimation in bacteria, and a method to provide an estimation of the number of human genes.

Acknowledgements

We would like to thank Oliver Clay for useful discussions. We also thank Giorgio Bernardi for his kind invitation to attend the 5th Anton Dohrn Workshop held in Ischia. This work is partially supported by grant BIO99-0651-CO2-01 from the Spanish Government. The help of David Nesbitt with the English version of the manuscript is appreciated.

References

Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J.L., 2002. Study of statistical correlations in DNA sequences. Gene 300, 105–115.
Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241, 3–17.
Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.
Boldögkoi, Z., Murvai, J., Fodor, I., 1996. G and C accumulation at silent positions of codons produces additional ORFs. Trends Genet. 11, 125–126.
Burset, M., Guigó, R., 1996. Evaluation of gene structure prediction programs. Genomics 34, 353–367.
Carpena, P., Bernaola-Galván, P., Román-Roldán, R., Oliver, J.L., 2002. On the number of human genes (in preparation).
Clay, O., Bernardi, G., 2001. Compositional heterogeneity within and among isochores and mammalian genomes (I and II). Gene 276, 15–31.
Devos, D., Valencia, A., 2001. Intrinsic errors in genome annotation. Trends Genet. 17, 429–431.
Guigó, R., 1997. Computational gene identification – an open problem. Comput. Chem. 21, 215–222.
Guigó, R., Agarwal, P., Abril, J.F., Burset, M., Fickett, J.W., 2000. An assessment of gene prediction accuracy in large DNA-sequences. Genome Res. 10, 1631–1642.
Hollom, T., 2001. Human genes: how many? Consolidation of information suggests more than 70,000. Scientist 15, 1.
Lander, E.W., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Macaya, G., Thiery, J.P., Bernardi, G., 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol. 108, 237–254.
Merino, E., Valvas, P., Puente, J.L., Bolivar, F., 1994. Antisense overlapping open reading frames in genes from bacteria to man. Nucleic Acids Res. 22, 1903–1908.
Oliver, J.L., Marín, A., 1996. A relationship between GC content and coding-sequence length. J. Mol. Evol. 43 (3), 216–223.
Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.
Oliver, J.L., Carpena, P., Román-Rodán, R., Mata-Balaguer, T., Mejías-

Romero, A., Hackemberg, M., Bernaola-Galván, P., 2002. Isochore chromosome maps of long human contigs. Gene 300, 117–127.

Senapathy, P., 1986. Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications. Proc. Natl. Acad. Sci. USA 83, 2133–2137.

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., Krogh, A., 2001. On the total number of genes and their length distribution in microbial genomes. Trends Genet. 17, 425–428.

Venter, J.C., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.