Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Physica A 348 (2005) 339–348
PHYSICA A
www.elsevier.com/locate/physa

# Exploring nonlinearity to identify genes and intergenic regions in genomes

Anastasios A. Tsonis[a,*], Panagiotis A. Tsonis[b]

[a]*Department of Mathematical Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201-0413, USA*
[b]*Department of Biology, University of Dayton, Dayton, OH 45469, USA*

Received 3 June 2004
Available online 13 November 2004

**Abstract**

In a previous paper (Physica A 312 (2002) 458) we presented an analysis, which established that DNA sequences exhibit significant nonlinear structure. More specifically, we found that both coding and non-coding sequences exhibit nonlinearity, but coding sequences are more nonlinear than non-coding sequences. Here, we extend this result and we propose an approach which explores nonlinearity to predict genes and intergenic regions in sequenced genomes. Our results compare favorably to other commonly used approaches in this problem.


*PACS:* 87.10

*Keywords:* DNA complexity; Linearity; Nonlinearity; Gene prediction

## 1. Introduction

With advances such as the complete human genome sequencing, the rat genome sequencing, and other major sequencing projects, a vast amount of data will be available to analyze for many decades to come. One of the central challenges for the

*Corresponding author. Tel.: +1 414 229 5373; fax: +1 414 229 4907.
*E-mail address:* aatsonis@uwm.edu (A.A. Tsonis).

0378-4371/$ - see front matter 
doi:10.1016/j.physa.2004.10.002

newly sequenced DNA is that of identifying the genes, the proteins they produce, and the functions these proteins perform. There are several aspects of the gene identification or prediction problem. The most common practice is to find the exons and introns and from there (through similarity searches) to predict proteins. In the last decade or so, several such approaches have been developed to address this aspect of the gene finding problem. Promising approaches include Glimmer [1,2] (an interpolated Markov model used for prokaryotic gene finding), Genscan [3] and Genie [4] (hidden Markov models used for eukaryotic gene finding), and Geneid [5]. Other approaches based on neural networks, decision trees and rule-based systems have also been suggested [6]. Most of these approaches are based on linear models (Markov processes are linear in character). Another aspect of the gene finding problem is to identify the regions in a genome sequence that correspond to individual genes and to separate them from intergenic regions. The above mentioned approaches can do this (by considering homologies, start and termination codons, etc.), but in genomic sequences (such as in the human chromosome 22) where there is a lot of intergenic DNA they tend to predict more genes than actual (because in those regions there is a good chance to find open reading frames (ORFs), which are not part of a true gene). In this paper, we will address both these aspects of the gene finding problem. More specifically, we will explore nonlinearities in the structure of DNA sequences to develop a new approach to identify genes and intergenic regions in genome sequences. Here, for the purpose of our analysis we will refer to a "gene" as a unit in a genome sequence made up of a 5′ UTR, coding regions, introns, a 3′ UTR, and other regulatory sites (such as promoters, start and stop codons, and the polyadenylation signal AATAAA), which may be regulated to code for a protein.

The nonlinear and linear component in DNA sequences was investigated in detail in Ref. [7]. The main idea is as follows: Given a sequence, $w(t)$, we consider an output $\tau$ steps ahead, $w(t+\tau)$, and its relation to input values lagged at $l_1, l_2, \ldots, l_n$, $w(t-l_1)$, $w(t-l_2) \ldots, w(t-l_n)$. With one output and $n$ input values we form the vector $\mathbf{Z}(\tau) = (\mathbf{X},\mathbf{Y}) = [w(t-l_1), \ldots, w(t-l_n), w(t+\tau)]$ and estimate the mutual information, $I$, between the inputs $\mathbf{X} = [w(t-l_1), \ldots, w(t-l_n)]$ and the output $\mathbf{Y} = w(t+\tau)$.

$$I(\mathbf{X},\mathbf{Y}) = \int_{R^d} p_{x,y} \ln \frac{p_{x,y}(\mathbf{x},\mathbf{y})}{p_x(\mathbf{x})p_y(\mathbf{y})}\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},$$

where bold characters indicate vectors, $x$ and $y$ are the values that the variables $X$ and $Y$ take, $p_{x,y}(\mathbf{x},\mathbf{y})$, $p_x(\mathbf{x})$, $p_y(\mathbf{y})$ are the joint and marginal probability distributions, and $d$ is the dimension of the space spanned by vectors $\mathbf{X}$ and $\mathbf{Y}$. The algorithm for the estimation of the mutual information for continuous real variables is described in Refs. [8,9]. Extension of this algorithm to discrete variables such as DNA sequences is described in Refs. [7,10]. This number, $0 \leqslant I \leqslant \infty$, plays a central role in information theory. It is usually normalized between zero and one using the transformation

$$\rho = \sqrt{1 - \mathrm{e}^{-2I}}.$$

The measure $\rho$ captures both the *linear* and *nonlinear* dependence between **X** and **Y** and it often interpreted as the predictability of **Y** by **X**. This measure is based on the probability distributions underlying the data and it does not depend on the particular model used to predict **Y** from **X**. The reason for choosing the above transformation is that $\rho$ reduces to some well-known measure of *linear* dependence when $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$ is a $d$-dimensional Gaussian random vector. In this case, the mutual information takes the form [8]

$$I(\mathbf{X},\mathbf{Y}) = \frac{1}{2}\ln\frac{\det\Sigma_{xx}\det\Sigma_{yy}}{\det\Sigma},$$

where $\Sigma$, $\Sigma_{xx}$ and $\Sigma_{yy}$ are the $d\times d$, $n\times n$ and $m\times m$ variance–covariance matrices of **Z**, **X**, and **Y**, respectively. In this case, it can be shown that the mutual information depends only on the coefficients of linear correlation [7,8]. It follows that the *linear* predictability of **Y** by **X** is given by

$$\lambda(\mathbf{X},\mathbf{Y}) = \sqrt{1-\frac{\det\Sigma}{\det\Sigma_{xx}\det\Sigma_{yy}}}.$$

One may think of $\rho$ as $\lambda+$nonlinearity. Thus, if a sequence exhibits nonlinearity, then $\rho>\lambda$. Tsonis et al. [7] found that in both coding and non-coding sequences the difference $\rho-\lambda$ is significantly greater than zero. They also found that because the estimation of nonlinear measures requires more data than the estimation of linear measures, if the data are limited (and this can happen either for a large number of inputs and/or for short genes), then $\rho$ is overestimated in relation to $\lambda$. This introduces a bias, $\varepsilon$. By estimating this bias from surrogate data they found that for $\tau = 50$ the difference, $x$, between the average (over the 50 steps ahead) of $\rho-\lambda$ ($\langle\rho-\lambda\rangle$) and the average bias, $\langle\varepsilon\rangle$, is greater for coding regions than for non-coding regions. It thus appears that nonlinearity may be a property of DNA sequences and that the degree of nonlinearity may vary between various sequence types in DNA. The purpose of this paper is to explore this observation and to develop an alternative method to identify genes and their parts in a genome.

## 2. Results

To this end, we considered the first 70,000 bp of the human chromosome 12p13. The structure of this sequence has been experimentally determined [11]. This is a challenging segment to consider for gene prediction algorithms because the intergenic regions and the exons are rather short in length [11]. In this segment we have the following:

| | |
|---|---|
| bp 1–1523: | intergenic region |
| bp 1524–32821: | gene CD4 |
| bp 32822–33808: | intergenic region |
| bp 33809–39430: | gene A (this gene has alternate spliced products) |
| bp 39431–41837: | intergenic region |

| | |
|---|---|
| bp 41838–51856: | gene B |
| bp 51857–52220: | intergenic region |
| bp 52221–59402: | gene GNB3 |
| bp 59403–60837: | intergenic region |
| bp 60838–63399: | gene C8 (complement) |
| bp 63400–64137: | intergenic region |
| bp 64138–70000: | first 5862 bp of gene ISOT |

While developing an approach to identify genes and parts of the genes in a genome, we have to keep in mind that some kind of a "scanning" procedure will be involved. If this procedure is computational, the "scanning" device could be a window of a pre-defined length, which considers the information in that window and then applies a mathematical operation. In this spirit, we considered non-overlapping intervals of 500 bp (i.e., a total of 140 intervals) and for each one we calculated for one input ($w(t-1)$) the average $\rho-\lambda$ over 50 steps ahead. Then for each of the 140 intervals windows we calculated $\langle \rho-\lambda \rangle$ and $\langle \varepsilon \rangle$ and estimated $x$. The corresponding 140 values of $x$ can be plotted as a histogram (top of Fig. 1) or as a function of the 500-bp interval (bottom of Fig. 1). Fig. 1 (top) indicates that the distribution of $x$ is tri-modal. The three modes are quite distinct and separated by (approximately) the values 0.04 and 0.11. We will call these values the separatices. Using gene and exon information for this particular segment from NCBI (second column, Table 1), we can compare the results in Fig. 1 with the actual beginning and end of the genes, and with the location of exons and intergenic regions. We observe the following:

(1) All values of $x$ greater than approximately 0.11 correspond to exons. For this particular segment, not even one intron corresponds to a value of $x>0.11$. This indicates that the method does not predict wrong exons often. Each such value, however, may not necessarily represent one exon only. Because in this example the exons are rather small and close to each other, a given $x>0.11$ may correspond to more than one exons. For example, in gene CD4 (between the first two arrows) the value $x=0.15$ in the 25th 500-bp interval (in the range of 12,000–12,500 bp) corresponds to exons 1 and 2. Thus, only the lower bound of the actual exons can be predicted with this approach.

(2) Values of $x$ between 0.04 and 0.11 correspond to introns and all other gene parts (UTRs, promoters, start, stop, etc.) which do not code and which may be too short to be clearly differentiated from introns. *We will refer to all these as non-coding regions*. Of course, once these regions are known they can be identified through consensus sequence, for example, splice sites for introns, conserved elements such as TATA box in promoters, etc. Comparison between Table 1 and Fig. 1 (bottom) indicates that only in six cases exons are associated with a value of $x<0.11$. These cases correspond to very short exons (typically, smaller than 100 bp) and as such their signature in a non-coding-dominated 500-bp interval is diminished. They are indicated by the letter M on the NCBI column in Table 1 and indicate a missing exon. For this example, and considering that there are 48 actual exons in the segment, we estimate that at the exon level the sensitivity (1−[# of missing exons/#

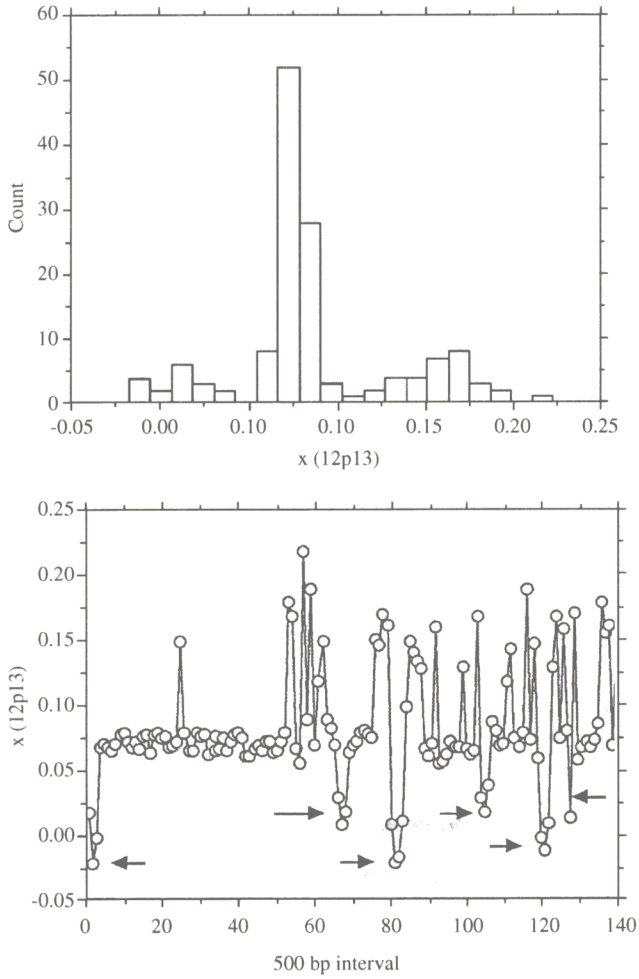


Fig. 1. (Top) Frequency distribution of the value of $x$ for the segment of human chromosome 12p13. Each value corresponds to a 500-bp interval. The distribution is clearly tri-modal. The first mode corresponds to intergenic regions, the second to introns and all other gene parts (UTRs, promoters, start, stop, etc.), which do not code, and the third to coding regions. We refer to the second category as non-coding regions. (Bottom) The value of $x$ along the segment. Since the total length of the sequence is 70000 bp there are 140 non-overlapping 500-bp intervals. As is explained in the text the intergenic regions are clearly identified. Accordingly, the number of genes in the sequence (six) is correctly predicted.

actual exons]) and specificity (1−[# of wrong exons/# of predicted exons]) of an approach which predicts exons for $x>0.11$, will be 87.5% and 100%, respectively.

(3) Values of $x$ less than 0.04 correspond to intergenic regions (indicated by the arrows). The fact that intergenic regions are characterized by very small $x$ values is not surprising. Intergenic regions include many repeat regions, which makes them strongly linear. In these cases $\rho \approx \lambda$ and the difference between $\rho$ and $\lambda$ is not as

Table 1
Gene and exon information for the 12p13 segment

| | NCBI | Geneid |
|---|---|---|
| CD4 (9) | | 1531...1650 (**F**) |
| | 12150...12199 | 12108...12199 |
| | 12319...12483 | 12319...12483 |
| | 26154...26312 | 26154...26312 |
| | 26771...27004 | 26771...27004 |
| | 28068...28415 | 28068...28415 |
| | 29142...29342 | 29142...29342 |
| | 30433...30554 | 30433...30554 |
| | 30859...30926 | 30859...30930 |
| | 31311...31341(**M**) | (**M**) |
| "A-1" | 35761...35775(**M**) | (**M**) |
| | | 35911...36777(**F**) |
| | 37495...37684 | 37495...37684 |
| | 38207...38364 | 38207...38364 |
| A (4) | 38664...39215 | 38664...38865 |
| | | 39204...39262 |
| "A-2" | 35911...36777 | |
| | 37495...37684 | |
| | 38207...38364 | |
| | 38664...39215 | |
| B (14) | | 40452...40667(**F**) |
| | 41926...42021 | 41869...42021 |
| | 42363...42564 | 42363...42564 |
| | 42649...42780 | (**M**) |
| | 43226...43362 | 43226...43362 |
| | 43835...43924 | 43835...43924 |
| | 45602...45654 | (**M**) |
| | 45752...45819 | (**M**) |
| | 45935...46059 | 45895...46059 |
| | 48995...49096(**M**) | 48995...49096 |
| | 49433...49583 | 49433...49583 |
| | 49742...49859 | 49742...49859 |
| | 49968...50043 | 49968...50043 |
| | 51010...51150 | 51010...51150 |
| | 51307...51471 | (**M**) |
| | | 53596...53643 |
| | | 54980...55086 |
| | | 55184...55247 |
| | | 55373...55535 |
| | | 55642...55708 |
| | | 55787...55988 |
| | | 57596...57812 |
| | | 58802...58908 |
| GNB3 (9) | 53298...53354(**M**) | (**M**) |
| | 53596...53634(**M**) | |
| | 54980...55086(**M**) | |
| | 55184...55247 | |

Table 1 (*continued*)

| | NCBI | Geneid |
|---|---|---|
| | 55373...55535 | |
| | 55642...55708 | |
| | 55787...55988 | |
| | 57596...57812 | |
| | 58802...58908 | |
| C8 (5) | 61053...61208 | 61053...61208 |
| | 61333...61439 | 61333...61439 |
| | 61575...61868 | 61575...61868 |
| | 62477...62606 | 62477...62606 |
| | 62843...62962 | 62843...62962 |
| ISOT (7) | 64190...64300 | 64190...64300 |
| | 67411...67536 | 67411...67536 |
| | 67765...67831 | 67765...67831 |
| | 68027...68160 | 68027...68160 |
| | 68315...68460 | 68315...68460 |
| | 68717...68901 | 68717...68901 |
| | 69639...69733 | 69639...69733 |

The first column indicates the name of gene and the number of its exons in the segment 12p13 of the human chromosome 12. The second column provides exon information from NCBI. In this column **M** indicates that this exon was missed by our approach. The third column shows prediction results from Geneid. Here **M** again indicates that this exon was missed by Geneid and **F** indicates that Geneid falsely predicted an exon.

significant. Exiting an intergenic sequence and entering a gene is always associated with a sharp increase in $x$, whereas exiting a gene and entering an intergenic region is associated with a decrease in $x$. Accordingly, all six genes can be identified by this procedure. Note that because of the formulation of $\rho$ and $\lambda$ our approach cannot distinguish a gene from a complement gene. The estimation of mutual information and of the covariance matrices does not depend on the transformation $A \rightarrow T$ and $G \rightarrow C$ or on the direction we read the information in the window.

In summary, although our approach may only predict the lower bound of the actual exons, its strength lies in: (1) its ability to predict very accurately the number of genes in a genome, (2) its ability to indicate that exon(s) exist in a given 500-bp interval (provided they are not too short), and (3) that it does not tend to predict wrong exons.

In order to appreciate our results better, we compared our method to other well-known approaches. Fig. 2 shows prediction results obtained by applying Genscan [3] and Geneid [5] to this segment. Both these approaches were able to find five genes (genes B and GNB3 were predicted as one gene). Inside these five genes several exons are also predicted. Note that both approaches predict more or less the same exons. The exon predictions from Geneid are shown in the third column of Table 1. As before the letter M identifies missing exons and the letter F identifies wrongly predicted exons (false alarms). Based on these results, we estimate that for Geneid

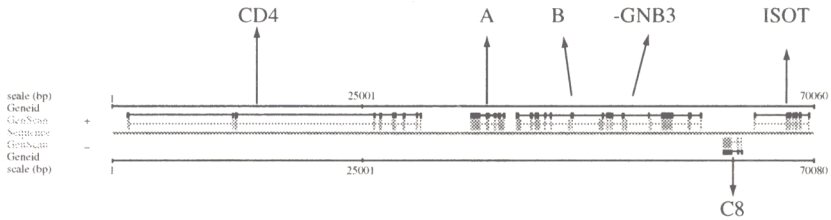


Fig. 2. Prediction of genes and exons for the same segment used in Fig. 1, but using Genscan and Geneid. Both approaches predict five out of six genes.

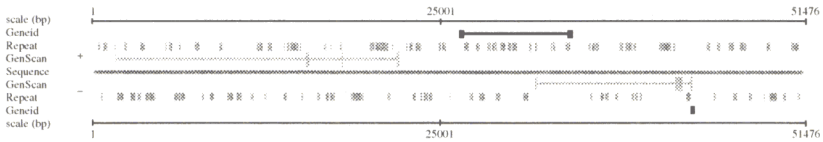


Fig. 3. Same as Fig. 2 but for the 51476-bp long sequence HS503F6 from clone CTA-503F6 of human chromosome 22q11.2-12.1.

and Genscan the sensitivity and specificity at the exon level is 85% and 93% respectively. It would thus appear that all approaches are quite comparable and successful. Note that Genscan performs very similar to Geneid. However, as will show next, our approach not only can be used in conjunction with Genscan or Geneid to improve predictions, but in certain cases it will be superior.

As we mentioned above one of the strengths of our method lies in its ability to indicate that exon(s) exist in a given 500-bp interval. Thus, it could be used to compliment and improve predictions of the actual number of exons obtained by the popular Genscan or Geneid or other approaches. For example, in the above analysis and comparison, in 4 out of the 7 Geneid M cases, our approach indicates that exons exist in the corresponding interval. A correction based on that would increase Geneid's sensitivity to 94%. This is approximately a 10% improvement.

Most important, is the ability of our approach to delineate intergenic regions clearly. One of the most serious problems with Genscan and similar approaches is that they tend to over-predict the number of genes in cases where large intergenic regions are present as, for example, in the human chromosome 22 [12,13]. To demonstrate this, we considered the 51476-bp long sequence HS503F6 from clone CTA-503F6 of chromosome 22q11.2–12.1. Fig. 3 is similar to Fig. 2 but for the chromosome 22 sequence. We observe that Genscan predicts two rather long genes (one is a complement gene) and Geneid predicts one gene. According to the information at NCBI (see also Ref. [14]) this sequence contains one putative gene in the range of 33,624–34,536 bp. More specifically, there is an exon in the range 33,624–33,770 bp, an intron in the range 33,771–34,378 bp, and an exon in the range 34,379–34,536 bp. This is in the region where Genscan predicts the complement gene

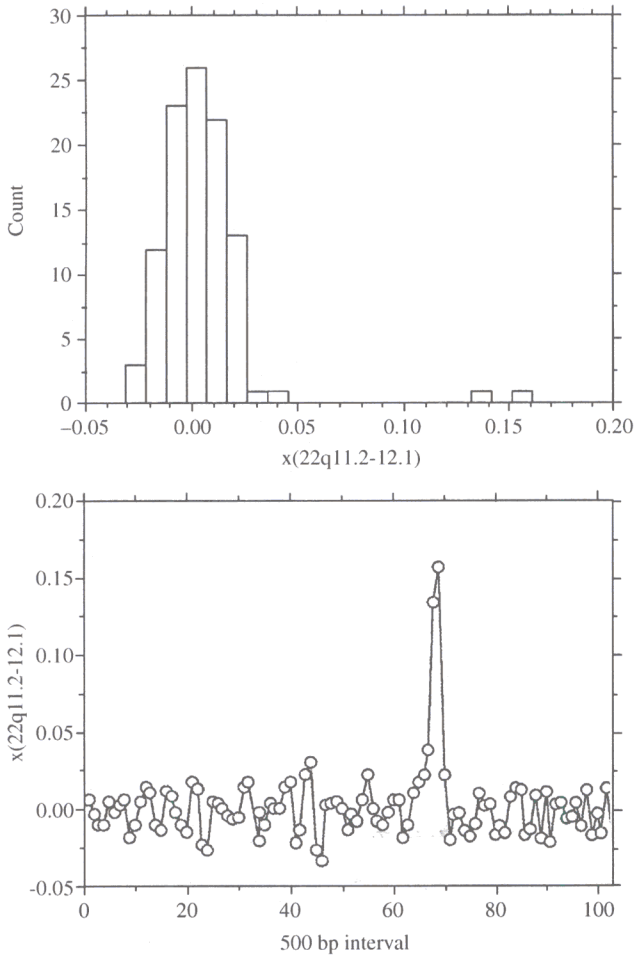


Fig. 4. Same as Fig. 1 but for the 51476-bp long sequence HS503F6 from clone CTA-503F6 of human chromosome 22q11.2-12.1. According to this figure one short gene is present, which according to information from NCBI is correctly identified.

but Genscan predicts a much longer gene (of the order of 10,000 bp). Geneid misses the suspected gene all together. Fig. 4 is similar to Fig. 1 and shows the results when the approach discussed above is applied blindly to this sequence. Accordingly, if $x$ is less than 0.4 the corresponding region is an intergenic region and if it is greater than 0.4 it indicates a gene. Furthermore, an $x$ greater than 0.11 indicates the presence of exons in the gene. Remarkably, our approach predicts that most of the segment is intergenic region except for the region where the suspected gene exists. This is an indication that the separatices may be rather robust. Clearly, our preliminary results demonstrate that the proposed approach can significantly improve the prediction of the total number of genes in a genome and that, once a gene is identified, it can

complement other approaches to further improve the prediction of exons and other regions in the predicted genes.

## 3. Conclusions

We have presented an approach, which explores the nonlinear structure of DNA sequences to identify genes and intergenic regions in sequenced genomes. We find that our approach is performing as good as other commonly used approaches when it comes to predicting the number of genes, but it outperforms them when it comes to identifying intergenic regions. In addition, we find that combining our approach with other approaches can improve exon prediction. We hope that our results will promote nonlinear approaches in gene finding projects.

## Acknowledgements

We would like to thank Victor Ruotti for his help with Figs. 2 and 3.

## References

[1] F. Jelineck, R.l. Mercer, Interpolated estimation of Markov source parameters from sparse data, in: E.S. Gelsema, L.N. Kanal (Eds.), Pattern Recognition in Practice, Elsevier/North-Holland, New York, 1980, pp. 381–397.
[2] S. Salzberg, et al., Microbial gene identification using interpolated Markov models, Nucleic Acids Res. 26 (1998) 544–548.
[3] C. Burge, S. Karlin, Prediction of complete gene structure in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[4] M. Reese, et al., Improved splice site detection in Genie, J. Comput. Biol. 4 (1997) 311–323.
[5] R. Guigo, Assembling genes from predicted exons in linear time with dynamic programming, J. Comput. Biol. 5 (1998) 681–702.
[6] S. Needleman, C. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[7] A.A. Tsonis, F.L. Heller, P.A. Tsonis, Probing the linearity and nonlinearity in DNA sequences, Physica A 312 (2002) 458–468.
[8] G.A. Darbellay, Predictability: an information-theoretic perspective, in: A. Prochazka, et al. (Eds.), Signal Analysis and Prediction, Birkhauser, Boston, 1998, pp. 249–262.
[9] G.A. Darbellay, I. Vajda, Estimation of the information by adaptive partitioning of the observation space, IEEE Trans. Inform. Theory 45 (1999) 1315–1321.
[10] I. Grosse, H. Herzel, S.V. Buldyrev, H.E. Stanley, Species independence of mutual information, Phys. Rev. E 61 (2000) 5624–5629.
[11] M.A. Ansari-Lari, Y. Shen, D.M. Muzny, W. Lee, R.A. Gibbs, Large-scale sequencing in human chromosome 12p13: experimental and computational gene structure determination, Genome Res. 6 (1996) 314–326.
[12] C.B. Burge, Personal communication.
[13] M. Das, C.B. Burge, E. Park, J. Colinas, J. Pelletier, Assessment of the total number of human transcription units, Genomics 77 (2001) 71–78.
[14] M. Swann, 1999, Direct submission, Sanger Centre, Hinxton, UK (humquery@sanger.ac.uk).