2011

The Analysis of Pattern Change in Intron Sequences

Jinda Kongcharoen
University of Wollongong

Yan-Xia Lin
University of Wollongong, yanxia@uow.edu.au

Rachel Caldwell
University of Wollongong

Yiren Yang
University of Wollongong

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

05-11

The Analysis of Pattern Change in Intron Sequences

Jinda Kongcharoen, Yan-Xia Lin, Rachel Caldwell, Yiren Yang and Ren Zhang

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

The Analysis of Pattern Change in Intron Sequences

Jinda Kongcharoen
Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong, NSW, Australia
jk117@uowmail.edu.au

Yan-Xia Lin
Centre for Statistical and Survey Methodology
School of Mathematics and Applied Statistics
University of Wollongong, NSW,
Australia yanxia@uow.edu.au

Rachel Caldwell
School of Biological Sciences
University of Wollongong, NSW, Australia
rac34@uowmail.edu.au

Yiren Yang
School of Biological Sciences
University of Wollongong, NSW, Australia
yy083@uowmail.edu.au

Ren Zhang
School of Biological Sciences
University of Wollongong, NSW, Australia
rzhang@uow.edu.au

Abstract— **The Generalized Bernoulli Modeling approach is used to analyze the pattern change in intron sequences of a model plant species *Arabidopsis thaliana*. The influence of the intron length and the number of GC on the intron sequence pattern changes is examined. Two other gene properties, the gene expression level and the protein function encoded are also assessed. Among the random sampled intron sequences, 10.71% of them have been identified to have sequence pattern change. Our study shows that the number of GC and the intron length significantly influence the intron pattern change while the gene expression level and the protein function have little effect. Our results show that for *Arabidopsis thaliana*, the shorter intron with more number of GC might have a higher chance to have pattern changes detected on its sequence and this piece of information could be used for checking whether the intron is functional introns. This study may be benefit to the further study on functions of intron.**

Keywords- pattern change; intron sequence; intron length; number of GC, GC content

I. INTRODUCTION

As the field of genomics is growing rapidly due to the increase in data availability of full genomes, more research attention is now focusing on studying the non-coding sequences [1, 2]. The up- and down-stream non-coding sequences are known to play a major role in the regulation of gene transcription and translation [3]. Introns, known as the non-coding regions within genes of eukaryotes, were once considered "junk" genes [4]. However, increasing research has suggested that certain introns are also involved in gene expression and regulation [5]. Although studies have shown that certain introns could promote expression [6, 7] or play important roles in transcription processes [5], the specific function of introns remains unknown. The accessibility of database of all intron sequences for certain species like *Arabidopsis thaliana*, now it carries out systematic analyses, but it is critical to develop a method to extract "functional introns", based on our current knowledge on gene structures. Studies have shown that GC-content and length of introns are important in gene regulation or expression [8] and are significantly correlated [4]. Understanding the influence and role of introns, in association with gene expression level has been a widely debatable subject. Many research studies [9-12] have explored the function of introns, in particular the length of introns in relation to gene expression level within the coding sequences (CDS) and untranslated regions (UTRs).

Statistical analysis methods can be used to examine the function and relevance of introns. DNA sequences can be modeled through a generalized Bernoulli process. The concept of "segmentation" in DNA sequences is equivalent to the concept of "stationary subsequence" in a stochastic process [13]. A position in a DNA sequence, which is used to link two consecutive segments, is called a "change point" in this paper. In terms of the concept of "stationary", a DNA sequence with sequence structure change/pattern change is equivalent to having at least one change point in the DNA sequence. Identifying the pattern changes in DNA sequences is significantly important in terms of discovering the genome functional components, understanding relevant evolutionary processes and describing genome architecture [14].

The use of statistics to detect change points in DNA sequences has been applied since the late 1980's [15 - 18]. Many different inference methods have been developed for the purpose of detecting change points in DNA sequences. They include Sequential Importance Sampling [14], Hidden Markov Model [19], Bayesian Hidden Markov Model [20, 21], the Cross-Entropy Method [22] and Generalized Bernoulli Modeling approach [13]. Previous researchers have studied the segmentation of DNA; however few have focused on intron segmentation studies. Many different methods can be applied for estimating change points in intron sequences. However, all methods have their own limitation to obtain accurate estimation of change points. The Maximum Likelihood method will often provide meaningless estimations [17]. The inference results on change points provided by Bayesian Hidden Markov models (Bayesian HMMs) and Hidden Markov models (HMMs) are very sensitive to prior knowledge about the location of the change points [19]. Furthermore, the methods of Bayesian

HMMs and HMM do not explicitly provide point estimations of change points. In this study, we employ the Generalized Bernoulli Modeling approach [13] and use this approach to estimate change points in introns. This method provides the visual-aid detection information on change points from the plots of the relevant associate processes. The information is then used to improve the Maximum Likelihood estimation of change points and provide more accurate estimation on change point in introns. In this study, we use *Arabidopsis thaliana* data as an example to demonstrate how to discover secret within introns through change point detection.

II. DATA COLLECTION AND ESTIMATION OF CHANGE POINT

The data used in this study and the change point estimation are obtained by following the workflow described in Fig. 1.

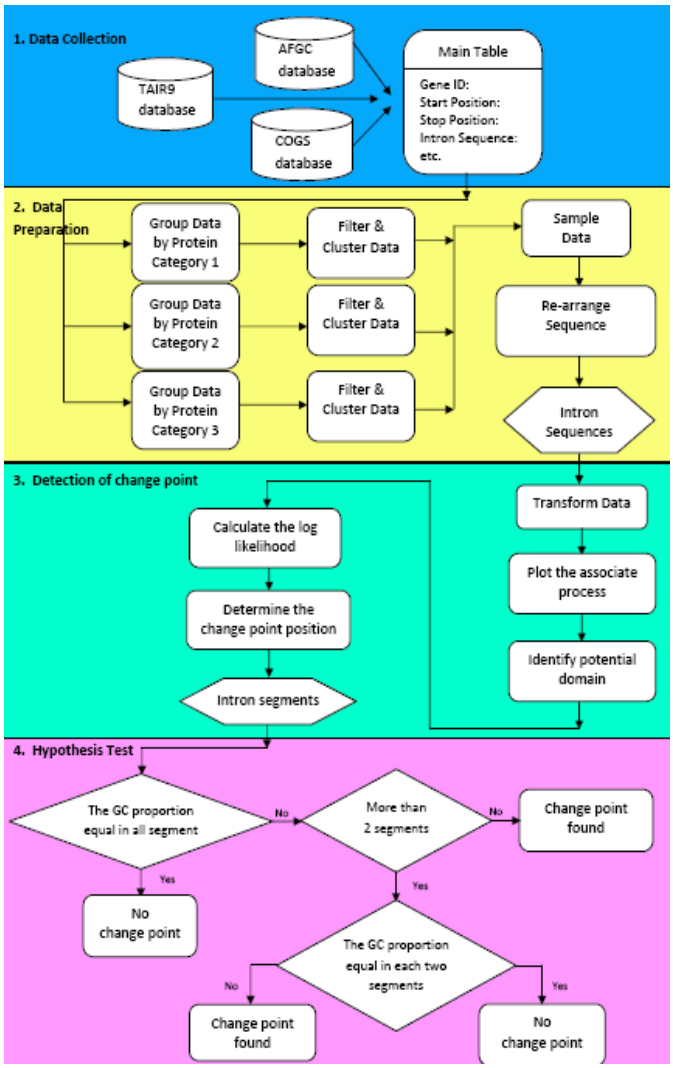

Figure 1. Data processing and data analysis workflow

A. Data collection

This study focuses on *Arabidopsis thaliana* which offers a compact genome and is widely used as one of the model organisms for studying plant genetics and development [23]. The intron sequence data was obtained from TAIR9 database (The Arabidopsis information resource) [24]. Protein functional category classification was obtained from the Clusters of Orthologous Groups of proteins database [25]. The gene expression data was collected from microarray data (The Arabidopsis Functional Genomics Consortium: AFGC) database [26]. By using standard identifiers, we link all intron sequences, protein function and gene expression data into one main table for analysis.

B. Data preparation

We used a proportional stratified random sampling design to randomly select intron sample for this study. A proportional stratified random sampling design is a process of grouping members of underlying population into relatively homogenous subgroups before sampling. The steps involved in constructing the sampling design are summarized as below:

- Under stratified sampling, heterogeneous genes are grouped by protein functional category (strata). Protein function is classified into three categories, Information storage and processing, Cellular processes and signaling, and Metabolism, by the functional classification of proteins in the Cluster of Orthologous Groups (COGs) database [25].
- Within each category, genes with the same locus tag ID are grouped into the same cluster. Also, genes with duplicate start and stop positions of the genome are filtered. The total numbers of gene clusters for Protein Categories 1, 2 and 3 are 903, 1478 and 1336 respectively.
- 50 samples of the gene clusters are randomly selected from the three defined strata in proportion to the number of gene clusters within each stratum. There are 12, 20 and 18 gene clusters in total for Protein Category 1, 2 and 3 respectively. The numbers of intron sequences in Protein Category 1, 2 and 3 are 44, 84 and 77 respectively. The total sampled introns are 205.
- All intron sequences in this study are in the forward direction.

C. Detection of change points

A DNA sequence can be expressed as a sequence of $Y_1, Y_2, \ldots, Y_n$ where Y_i takes one of DNA alphabet (A, C, G or T). Sometimes G-C base-pairs model, A-G base-pairs model or T-G base-pairs model are considered in DNA sequence analysis. For example, for G-C base-pairs model, bases G and C are classified into a same category. Those models can be modeled through generalized Bernoulli processes.

Definition: A process Y_t is called a generalized Bernoulli process, if for all $t > 0$, Y_t has Bernoulli distribution with mean $p_t > 0$.

In the definition of a generalized Bernoulli process, $\{Y_t\}$ are not necessarily mutually independent. However, to model G-C base-pairs in this paper, we adopted previously suggested method from the literature and accept that $\{Y_t\}$ are mutually independent [16, 27]. Therefore, to model a DNA sequence by a G-C base-pairs model, a generalized Bernoulli process Y_t is defined as follows: $Y_t = 1$ if the sequence at its t^{th} position is G or C; otherwise $Y_t = 0$ [16, 27].

Given a generalized Bernoulli process $Y_1, Y_2, \ldots, Y_n$, if there is an integer $\tau \in (1, T)$ such that for all $t < \tau$, $p_t = a$ and for all $t > \tau$, $p_t = b$ where $a \neq b$ then τ is called a change point in the sequence $Y_1, Y_2, \ldots, Y_T$.

To gain insight into the relationship among intron sequences, gene expression level and protein function**,** firstly we examine each sampled intron sequence to find out whether the intron sequence has pattern changes/change points or not. We use G-C model to model intron sequence. The process used for detecting change points involves four steps [13]. Step 1, for each intron sequence, define a sequence $y_1, y_2, \ldots, y_n$ such that, $y_i = 0$ if the i^{th} position of the intron sequence is A or T; $y_i = 1$, otherwise (C or G). Step 2, define an associate sequence from the original sequence and produce a plot of the associate process. Step 3, observe the pattern changes in the plot of the associate process and identify potential region(s) for change point(s) if there are any significant pattern changes appearing on the plot. Step 4, apply the Maximum likelihood method to the potential regions and estimate the change point(s).

D. Hypothesis test

After change points in an intron sequence are detected, the intron sequence will be divided into segments by the positions of those change points. Each segment is presumed stationary; its mean is estimated by the proportion of GC in the segment. For each segment, the GC proportion was calculated. To further check the accuracy of the estimates of change points, we use the z-test to test if the GC proportions are significantly different for any two consecutive segments which share a change point as one end of the two segments. The significance level used in the test is 0.05. The null hypothesis is that there is no difference between the GC proportions between the two tested segments. If test statistic z <–1.96 or z >1.96, the null hypothesis will be rejected. It indicates that there is evidence to support that the means of the two tested segments are different and therefore it further confirms the shared end of the two segments is a change point. If the null hypothesis is accepted, the two tested segments will be combined together and form a longer segment. In addition, to test if more than two segments (more than one change point is found) have the same mean or not, a Chi-square test will be used.

Using the intron sequence AT1G06490 in Protein Category 2 (consisting of 73 bp) as an example, we illustrate the process above.

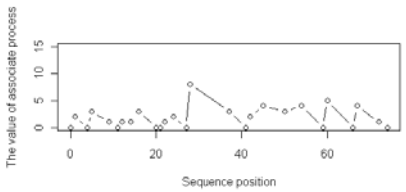

Figure 2. The plot of associate process given by intron AT1G06490

Given the AT1G06490 sequence, the plot of associate process is shown in Fig. 2. From the plot, the plot pattern change is easily identifiable. It suggests there is a change point in the sequence and a potential domain (20, 40) for the change point is determined. The sequences from 1 to 20 bp and from 40 to 73 bp are considered to be stable. The means of sequences are calculated respectively and then will be used in the process of estimating the possible change point between positions 20 and 40 [13]. Within the potential domain, we calculate the value of a log likelihood function defined in [13] and use the Maximum likelihood method to estimate the change point position. The estimated position of the change point is 28 where the log likelihood function takes the maximum. The scatter plot of the log likelihood function is shown in Fig. 3. After the change point is estimated, we calculate the means of the segments from 1 bp to 28 bp and 29 bp to 73 bp respectively. The results are 0.5000 and 0.2444. To further check if we are able to reject that there is no significant difference between the means of these two segments at significance level 0.05, a z-test is carried out. The value of test statistic z is 2.2374 indicating there is evidence to support that the two means are significantly different. Therefore the position of change point in this sequence is 28.

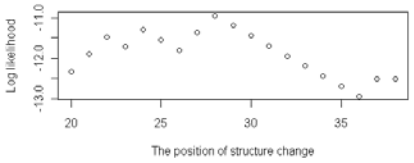

Figure 3. The scatter plot for log likelihood function

III. RESULTS

A. Compairson of the intron length and the number of GC

We divide 205 random sampled intron sequences into 2 groups: 22 intron sequences with pattern change (Group 1) and 183 intron sequences with no pattern change (Group 2). The median of the length of intron sequences in Group 1 is 89 bp and ranges from 65 – 1188 bp. The median of the length of intron sequences in Group 2 is 105 bp and ranges from 69 - 792 bp. Higher skewness (3.2291) can be seen in the length distribution of Group 1 compared to the length distribution of Group 2 (1.8881). Difference in kurtosis is also showed between these two groups, indicating the distribution of the intron length in Group 1 tends to have higher distinct peak near the mean. In addition, the median of the number of GC in Group 1 is 28 and ranges from 18 – 441; the median of the number of GC in Group 2 is 34 and

ranges from 17 – 275. Both medians of the length of intron and the number of GC in Group 2 are greater than those for Group 1. Applying Mann-Whitney U test to the data, there is evidence to support that the medians of the length of an intron sequence and the number of GC in Group 1 are less than those in Group 2. This suggests that the longer the intron is, the less the chance the intron has pattern change in its sequence.

B. Analysis of GC content within intron (without pattern change) in different protein functional groups

Those sequences that do not contain pattern change are also analyzed to test whether these intron sequences have the same probability structure or not. The mean of the GC number in each intron is the main measurement to test whether these intron sequences have the same probability structure or not. If two introns (without pattern change) have the same mean of GC number, these two introns can be accepted to have the same probability structure. Therefore, for each sampled gene, we used the Chi square test to test whether all the introns (without pattern change) in the gene have the same GC content. It turns out that the conclusion is held at significant level 0.05.

To test the GC content within different genes, consideration of the intron sequences (without pattern change) in all genes by protein function are also performed by using the Chi square test. The analysis indicates that there is evidence to support that all of the introns (without pattern change) in Protein Category 1 (p-value = 0.1459) and Protein Category 2 (p-value = 0.292) can be accepted with the same GC content, while no evidence to support that all of the introns (without pattern change) in Protein Category 3 (p-value < 0.000) have the same probability structure. The GC content of sampled intron sequences (without pattern change) in both Protein Categories 1 and 2 are stable and their pooled GC content can be accepted at 0.3210 and 0.3293 respectively. Previous research examining the GC content within introns in the *Arabidopsis* found it was 32%, confirming these results [28].

C. The relationship between the gene properties and intron sequence pattern change

Little is known about the impact of intron pattern change on the gene expression level and protein function. In this study, we investigate whether there is any connection between the gene expression level or protein function against intron pattern change in genes. A gene may contain one or more introns. If one of its introns has pattern change, we will define the gene as the gene with intron pattern change.

For each sampled intron sequence, it is classified into different groups based on gene expression level and the protein function. Regarding gene expression intensity, a gene with average gene expression intensity greater than 4775.683 (the median value of the whole average gene expression intensity) will be in the high gene expression level, otherwise in the low gene expression level. Our hypothesis is that a gene with higher average gene expression may have more chance to be a gene with intron pattern change. A Chi-square test shows that there is no sufficient evidence to support our hypothesis test.

When taking protein function into account, a Chi-square test also shows that there is no evidence to support that a gene belonging to a particular protein functional category will be more likely to have intron sequences pattern change**.** Our study also confirms in biology that the structure of intron sequences should be irrelevant to gene expression level and protein function.

D. Analysis of GC content and intron pattern changes

Within 205 sampled introns, 22 introns have been detected to have pattern change(s). Studying the GC content in each segment in each intron found that 18 introns out of 22, each of them has at least one intron segment with GC content greater than 0.36 (which is the mean GC content of *Arabidopsis thaliana* [29]). A z-test is conducted to determine if there is significant that more than 50% of intron sequences with pattern change have higher GC content in one of their intron segments. The test gives p-value 0.0028 and indicates that there is evidence to support more than 50% of intron sequences with pattern change have higher GC content in one of their intron segments.

E. The relationship between the probability of an intron having pattern change against the length of intron and the number of GC in the intron sequence

We applied the binary logistic regression analysis to model the relationship between the probability of an intron with pattern change against the length of intron as well as the number of GC in the intron. The best fit model is

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_1 length + \beta_2 GC + \beta_3 length \times GC$$

where $\pi(x)$ is the conditional probability of an intron sequence with pattern change; the estimations of $\beta_1, \beta_2, \beta_3$ are given by Table I

TABLE I. THE SUMMARY OF LOGISTIC REGRESSION ANALYSIS

Variables	**Coefficient**	**Standard error**	**t -value**
Length	-0.0181	0.0090	-2.010*
GC	0.1277	0.0319	4.002 *
Length*GC	-0.00008	0.00002	-3.732*****

* is significant at the 0.05 level.

The estimation of the conditional probability of an intron having pattern change is given by the following equation:

$$\pi(x) = \frac{e^{-0.0181(length)+0.1277(GC)-0.00008(length\times GC)}}{1+e^{-0.0181(length)+0.1277(GC)-0.00008(length\times GC)}} \quad (1)$$

Based on (1), we plot the probability of the pattern change in the intron sequence versus the intron length for each given numbers of GC (27, 34 and 71) (Fig. 4). The plot clearly shows the probability of the pattern change in the intron sequence decreases as the length of the intron sequence increases. It also shows that the higher number of GC in the intron, the higher probability of the intron will have pattern change. This might imply that the shorter length and the higher the number of GC in an intron might have correlation to the pattern change within intron sequences. This result of *Arabidopsis thaliana* is similar to the result of *Drosophila* that patterns of intron sequence is dependent on the intron length and the number of GC [30].

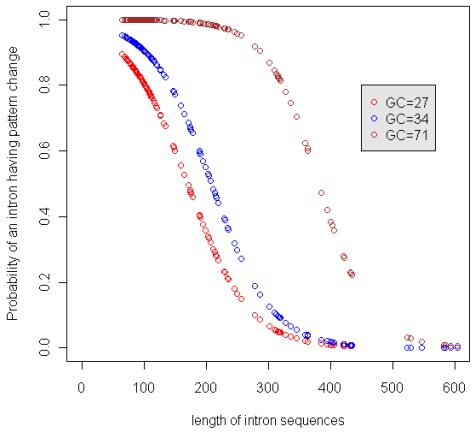

Figure 4. Probability of having pattern change in relation to the length of intron sequences given number of GC = 27, 34 and 71

IV. Discussion

Preliminary results using several statistical methods have shown that intron sequences that displayed pattern change have high values of skewness and kurtosis and experience shorter median intron lengths. Little investigation has been undertaken in the length distributions of these sequences, in particular in relation to pattern change. To extend on the preliminary results, this was investigated.

The relationship between the pattern change of intron sequences and the length and the number of GC was examined. Comparisons of length and the number of GC showed significant differences, therefore a logistic regression model was used to establish whether there was a relationship between the presence of pattern change and the length of intron sequences and the number of GC. In summary, it was shown that pattern change was associated with shorter lengths of intron sequences and the higher number of GC. What role does intron's have in relation to the organisation, distribution and functional significance of genes has been a continuous area of study. Organisation of the genome into harmonized segments may reveal the significance the length of the intron and GC content has on change points within a gene, and genome itself. Previous research conducted on the *Drosophila* has found that short introns make up only a small percentage of total intronic DNA in the genome, and found that intron length and sequence evolution is negatively correlated [31]. It appears that shorter introns are less likely to diverge, there is strong evidence evolutionary constraints occurs on longer introns [31]. This includes the first intron where the length is an important attribute [32]. Length is also important if genes are reliant on improving the rate of mRNA export, and research has shown that minimal introns has some influence on function by enhancing the export of the mRNA [33]. Short introns have also been related to the acquisition of stroma-targeting peptides in the flagellate *Euglena gracilis*. The splicing and organization of exons and introns has been an important factor in the acquisition of chloroplast targeting signals [34].

A key area for many biologists is to gain more understanding of introns, and determine whether or not they have any functional importance within the genome. Research conducted on the human genome as well as the *Drosophila* had identified distinct patterns, particularly in association with the length of and the GC content of introns. The *Drosophila* research [4] has found that longer introns and lower GC content tend to have a higher probability to be functional introns. However, other research has determined within the human and chimpanzee genomes [5, 35] that higher GC content and shorter introns offer a higher probability that the introns are functional. Using this information already gained by other research work, it may support what this research paper has also found, investigating pattern change. We can surmise for *Arabidopsis thaliana* that shorter introns and higher GC number leads to a higher probability of pattern change within the introns, therefore suggesting that these introns can be described as functional. This current research and previous research on introns has showed the importance of length distributions in relation to biological processes and may facilitate in the building of a model that could identify and describe functional introns within multiple organisms. The relationship between the level of gene expression and pattern change showed a weak relationship. Previous research has reported that there is strong correlation between gene expression and GC-content [36], however [37] re-examined these finding and found very weak correlations, on individual genes, not on groups of genes. This research has built a better understanding of how introns are constructed and influence the function and replication of proteins within the organisms' genome.

V. Conclusion

With the limited sample size, our study does not reveal relationships between the existence of pattern change and the gene expression level as well as the protein function. However, as far as all introns (without pattern change) are concerned, the same GC content in introns in the same protein category is found in both Information storage & processing or Cellular processes & signaling protein categories but not found in Metabolism category.

Currently there is no research on the relationship between GC content and intron function in any plants and there is still no knowledge about if the length or GC content will have impact on the function of intron in any plants either. There is still little understanding on plant's introns.

Our study in Arabidopsis thaliana indicates that the intron length and the number of GC are statistically significantly correlated to the pattern change in the intron sequences. It also shows that a short intron with more number of GC has higher chance to have pattern change in its sequence. Based on statistical data analysis, we found that the length of intron and the GC content of intron have significant impact on the probability of an intron having pattern changes in its sequence.

REFERENCES

[1] F. Mignone, C. Gissi, S. Liuni, and G. Pesole, "Untranslated regions of mRNAs," Genome Biology, vol. 3, pp. 0004.1-0004.10, 2002.

[2] R. Caldwell, Y.X. Lin, and R. Zhang, "Correlations of length distributions between non-coding and coding sequences of *Arabidopsis thaliana*," Bioinformatics and Biomedicine IEEE International conference, pp. 72-77, 2008.

[3] T. Tuller, E. Ruppin, and M. Kupiec, "Properties of untranslated regions of the *S. cerevisiae* genome," BMC Genomics, 2009, doi:10.1186/1471-2164-10-391.

[4] L. Zhu, Y. Zhang, W. Zhang, S. Yang, J.Q. Chen, and D. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," BMC Genomics, 2009, doi:10.1186/1471-2164-10-47.

[5] E. Gazave, T. Marqués-Bonet, O. Fernando, B. Charlesworth, and A. Navarro, "Patterns and rates of intron divergence between humans and chimpanzees", Genome Biology, 2007, doi:10.1186/gb-2007-8-2-r21.

[6] G.C.Curi, R.L. Chan, and D.H. Gonzalez, "The leader intron of Arabidopsis thaliana genes encoding cytochrome coxidase subunit 5c promotes high-level expression by increasing transcript abundance and translation efficiency", Journal of Experimental Botany, vol. 56, no. 419, pp. 2563 - 2571, 2005.

[7] S.R. Norris, S.E. Meyer, and J. Callis, "The intron of Arabidopsis thaliana polyubiquitin genes is conserved in location and is a quantitative determinant of chimeric gene expression", Plant Molecular Biology, vol. 21, no. 5, pp. 895-906, 1993.

[8] K.R. Kalari, M. Casavant, T.B. Bair, H.L. Keen, J.M. Comeron and T.L. Casavant et al. "First exons and introns - a survey of GC content and gene structure in the human genome," In Silico Biology, vol. 6, no.3, pp. 237-242, 2006.

[9] X. Hong, D.G. Scofield, and M. Lynch, "Intron size, abundance, and distribution within untranslated regions of genes," Molecular Biology Evolution, vol. 23, no. 12, pp. 2392-2404, 2006.

[10] J. Colinas, S.C. Schmidler, G. Bohrer, B. Iordanov, and P.N. Benfey, "Intergenic and genic sequence lengths have opposite relationships with respect to gene expression," PLos ONE, 2008, doi:10.1371/journal.pone.0003670.

[11] B.Y.W. Chung, C. Simons, A.E. Firth, C.M. Brown, and R.P. Hellens, "Effect of 5' UTR introns on gene expression in Arabidopsis thaliana," BMC Genomics, 2006, doi:10.1186/1471-2164-7-120.

[12] T.A. Hughes "Regulation of gene expression by alternative untranslated regions," Trends in Genetics, vol. 22, no. 3, pp. 119-122, 2006.

[13] Y.X. Lin, "Visually identifying potential domains for change points in generalized Bernoulli processes: an application to DNA segmental analysis," Statistics working paper series, Centre for Statistical and Survey Methodology, University of Wollongong, 2009.

[14] G.Yu Sofronov, G.E. Evans, J.M. Keith, and D.P. Kroese, "Identifying Change-points in Biological Sequences via Sequential Importance Sampling," Environmental Modeling & Assessment. vol. 14, no. 5, pp. 577-584, 2009.

[15] G.A. Churchill, "Stochastic models for heterogeneous DNA sequences," Bull. Math. Biol, vol. 51, pp. 79-94, 1989.

[16] J.V. Braun, and H.G. Muller, "Statistical methods for DNA sequence segmentation," Statistical Science, vol. 13, pp. 142–162, 1998.

[17] P.J. Avery, and D.A. Henderson, "Detecting a changed segment in DNA sequences," Journal Of the Royal Statistical Society Series C, vol. 48, no. 4, pp. 489–503, 1999.

[18] J.L. Oliver, R. Roman-Roldan, J. Perez, and P. Bernaola-Galvan, "Segment: identifying compositional domains in DNA sequences," Bioinformatics, vol. 15, no. 12, pp. 974–979, 1999.

[19] R.J. Boys, D.A. Henderson, D.J. Wilkinson, "Detecting homogeneous segments in DNA sequences by using Hidden Markov models," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 49, no. 2, pp. 269-285, 2000.

[20] D. Nur, D. Allingham, J. Rousseau, K. Mengersen and R. McVinish, "Bayesian hidden Markov model for DNA sequences segmentation: A prior sensitively analysis segments," Computational Statistics and data analysis, vol. 53, no. 5, pp. 1873-1882, 2009.

[21] R.J. Boys, D.A. Henderson, "A Bayesian approach to DNA sequence segmentation," Biometrics, vol. 60, no.3, pp. 573-588, 2004.

[22] G.E. Evans, G.Yu Sofronov, J.M. Keith, D.P. Kroese. "Estimating Change-Points in Biological Sequences via the Cross-Entropy Method", Annals of Operations Research, in press.

[23] S.M. Coelho, A.F. Peters, B. Charrier, D. Roze, C. Destombe and M. Valero et al. "Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms," Gene, vol. 406, pp. 152–170, 2007.

[24] The Arabidopsis information resource [http://www.arabidopsis.org/].

[25] The Clusters of Orthologous Groups of proteins database [http://www.ncbi.nlm.nih.gov/COG/].

[26] The Arabidopsis Functional Genomics Consortium database [ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/].

[27] Y. Fu, and R. Curnow, "Maximum likelihood estimation of multiple change points," Biometrika, vol. 77, no. 3, pp. 563-573, 1990.

[28] A.B. Rose, T. Elfersi, G. Parra, and I. Korfa, "Promoter-proximal introns in Arabidopsis thaliana are enriched in dispersed signals that elevate gene expression," The Plant Cell, vol. 20, pp. 543–551, 2008.

[29] The National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/].

[30] P.R. Haddrill, B. Charlesworth, D.L. Halligan, and P. Andolfatto, "Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biology, 2005, doi:10.1186/gb-2005-6-8-r67.

[31] P. Andolfatto, "Adaptive evolution of non-coding DNA in *Drosophila*," Nature, vol. 437, pp.1149-1152, 2005.

[32] G. Marais, P. Nouvellet, P. Keightley and B. Charlesworth, "Intron size and exon evolution in *Drosophila*," Genetics, vol. 170, pp. 481-485, 2005.

[33] J. Yu, Z. Yang, M. Kibukawa, M. Paddock, D. Passey, G.K. Wong, "Minimal introns are not junk," Genome Research, vol. 12, pp. 1185-1189, 2002.

[34] M.Vesteg, R.Vacula, J.M. Steiner, B. Mateášiková1, W. Löffelhardt, B. Brejová et al. "A Possible Role for Short introns in the Acquisition of Stroma-Targeting Peptides in the Flagellate *Euglena gracilis*," DNA Research, vol. 17, no. 4, pp. 223-231, 2010.

[35] R.Versteeg, B.D.C.v Schaik, M.F.v. Batenburg, M. Roos, R. Monajemi, H. Caron, , H.J. Bussemaker, and A.H.C.v. Kampen, "The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes," Genome Research, vol. 13, no. 9, pp. 1998-2004, 2003.

[36] M.J. Lercher, A.O. Urrutia, , A. Pavlicek, and L.D. Hurst," A unification of mosaic structures in the human genome", Human Molecular Genetics, vol. 12, pp. 2411–2415, 2003.

[37] M. Sémon, D. Mouchiroud and L. Duret " Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance," Human Molecular Genetics, vol. 14, no. 3, pp. 421-427, 2005.