International Journal of
Computer & Software Engineering

Research Article Open Access

Detection and Analysis of Pseudogenes in Non-coding DNA

Gabriella Trucco* **and Vittorio Cerioli**
Department of Computer Science, University of Milan, via Celoria, Milan, Italy

Abstract

It is well known that elements lying outside the coding regions of the human genome are involved in many human diseases. Therefore, the efforts to detect and characterize functional elements in the non-coding regions are rapidly increasing. Among many types of non-coding DNA, pseudogenes are sequences that share some similarities with their parental genes but have lost their ability to code for proteins. In this paper, we propose a methodology for detection and analysis of pseudogenes, based on transition probabilities of the nucleotides and their occurrences. The 1000 base pairs length downstream region of each potential pseudogene is analyzed in order to find a polyA tail and a polyadenylation signal. We implemented a Hidden Markov Model with the Viterbi algorithm to decode the upstream regions of the previously detected pseudogenes in order to search for CpG islands. In order to identify motif signals in the selected pseudogenes, we implemented the Gibbs sampling algorithm and we executed it on the flanking regions of some pseudogenes. Results demonstrate that the proposed methodology is an efficacious solution to detect new potential loci, especially when the query coverage of the alignment is shorter than the coding sequence. These loci can be classified to pseudogene fragments.

Publication History:

Received: March 16, 2021
Accepted: April 17, 2021
Published: April 19, 2021

Keywords:

Pseudogenes, CpG island, Alignment, Viterbi algorithm, Gibbs sampling

Background

The epigenome comprising different mechanisms e.g. DNA methylation, remodeling, histone tail modifications, chromatin micro RNAs and long non-coding RNAs, interact with environmental factors like nutrition, pathogens, climate to influence the expression profile of genes and the emergence of specific phenotypes. Multi-level interactions between the genome, epigenome and environmental factors might occur [47]. Completing the human genome reference sequence was a milestone in modern biology. It was quickly recognized that nearly 99% of the ~3.3 billion nucleotides that constitute the human genome does not code for proteins [1]. More recently, studies have discovered many loci that contribute to human diseases and susceptibility lying outside the protein coding regions [2-8]. These findings suggest that the non-coding regions of the human genome contain a plentiful and variegated set of functionally significant elements. There are several segments of non-coding regions including: non-coding RNA, cis- and trans-regulatory elements, introns, pseudogenes, telomeres, transposons and repeat sequences. These regions seem to be responsible for a varied number of diseases in humans and, therefore, understanding their roles in the genome is of utmost necessity [2-8]. Actually, the claim that a "biochemical function" can be assigned to 80% of the human genome, made by some authors, has aroused severe criticisms [9,10]. Nevertheless, the interest in identifying and characterizing all functional elements in the human genome is widespread and fast growing.

Pseudogenes

A *pseudogene* is a genomic DNA sequence that is closely related to a gene but has lost the capacity to produce a functional protein. The estimated number of pseudogenes in the human genome is comparable to that of protein coding genes (~20.000) [11,12]. Some pseudogenes are clearly non-functional gene relics [13]. Other pseudogenes, on the contrary, although not translated into proteins, are capable of influencing the activity of other genes by means of long non-coding RNA (lnc RNA) transcripts. In particular, it was observed that some pseudogene transcripts can regulate the expression of their parental genes through competition for cytoplasmic RNA-stabilizing factors or, the other way round, for trans-acting destabilizing factors [14,23]. Moreover, some pseudogenes produce antisense transcripts capable to hybridize with their parental transcripts and leading to suppression of their translation [15].

There are three types of pseudogenes. *Unitary pseudogenes* are formed when spontaneous mutations occur in a coding gene that lead to the loss of either coding or transcription potential (Figure 1A). Unitary pseudogenes are the rarest class of pseudogenes (~100 in humans) [16]. A second class of pseudogenes, the *duplicated pseudogenes*, is produced by genomic DNA duplication when it is performed incorrectly resulting in a non-functional pseudogene (Figure 1B). A duplicated pseudogene retains the basic structure of functional genes, for example, promoters and introns [17]. The third class, known as the *processed pseudogenes*, is formed when a mature mRNA is reverse transcribed and integrated into a new location in the host genome (Figure 1C). Because processed pseudogenes are produced from mature mRNA, they usually lack introns and a promoter but, sometimes, they maintain a residue of the RNA polyA tail [18], a stretch of RNA that contains only adenines. In eukaryotes, the presence of a polyA tail at the 3' extremity is a feature of a mature messenger RNA. Polyadenylation is the addition of a polyA tail to a 3' messenger RNA. Processed pseudogenes are the most abundant class in humans [12,19] and are formed from just 10% of the coding genes. The type of genes that produce processed pseudogenes are principally highly expressed genes such as genes for ribosomal proteins [19,20].

Pseudogenes are transcribed only if they are integrated close to a pre-existing promoter [12]. Estimates of the number of transcribed pseudogenes suggest that as many as one fifth of the pseudogenes may be transcribed into RNA [12], with processed pseudogenes tending to be transcribed more often than duplicated pseudogenes [21]. Interestingly, several pseudogenes exhibit tissue specific patterns of

***Corresponding Author:** Dr. Gabriella Trucco, Department of Computer Science, University of Milan, via Celoria, Milan, Italy; E-mail: gabriella.trucco@unimi.it

Citation: Trucco G, Cerioli V (2021) Detection and Analysis of Pseudogenes in Non-coding DNA. Int J Comput Softw Eng 6: 164. doi: https://doi.org/10.15344/2456-4451/2021/164

transcription [18]. Pseudogene RNA levels can also change during differentiation [22] and in diseases such as cancer [23,24] and diabetes [25]. Although pseudogenes are considered to be evolving neutrally, there are many evidences of evolutionary conservation [26,27]. Moreover, pseudogenes with higher levels of evolutionary constraints show greater tendency towards being transcribed [28]. The findings that some pseudogenes have evolved under positive selective pressure and their transcripts can occur in a dynamic and tissue specific manner suggest that they may have an important role.

Characterizing the pseudogenes and understanding their regulatory role is essential to discover the genetic background of many diseases and to identify new pharmacological treatments. Moreover, the correct identification of pseudogenes is important also for gene annotation. Indeed, the prevalence of pseudogenes in mammalian genomes can introduce artifacts in automatic gene annotation pipelines in which pseudogenes are often mistakenly annotated as genes [12,44]. This is due to the high sequence similarity of pseudogenes with their parental genes.

Aims of the research

Protein sequence similarity to parent gene is the main feature used to detect pseudogenes, because it is deemed the most sensitive indicator [30,31]. In spite of this, we developed a methodology that is based on raw nucleotide identity (DNA sequence similarity) with the coding sequence (CDS) of the corresponding gene and on its transition probabilities. The coding sequence is the portion of the gene that remains in the mature messenger RNA after the splicing and, therefore, it is the portion that is actually translated into protein. It is composed by the exons. Once identified a putative pseudogene, we analyzed both the upstream (before the pseudogene 5' extreme) and the downstream (beyond the pseudogene 3' extreme) regions in order to find out biologically interesting features. In particular, we searched for a CpG island and promoter signals in the upstream region. Dinucleotide clusters of CpG, or CpG islands, are present in the promoters and in the exons of ~40% of mammalian genes. However, the abundance of CpG dinucleotides in the human DNA is much lower than expected based on the CG content because CpG dinucleotides, outside the promoters and the exons, are largely methylated. The low occurrence of CpG islands outside these regions is due to the fact that methylated cytosines are mutational hotspots and, as a consequence, they are depleted by natural selection [45,46]. The downstream region, instead, was analyzed in order to detect the presence of a polyA tail and, when a polyA tail was found, a polyadenylation signal was searched for. The polyadenylation signal (typically AAUAAA) is a binding site on the messenger RNA where polyadenylation starts. Moreover, we implemented the Gibbs sampling algorithm with the aim of finding a common motif in the upstream regions containing a CpG island [32,33].

Main results

In this paper we considered 11 genes and searched for their processed pseudogenes. Five of these genes belong to the ribosomal protein family, which is the family with the highest number of processed pseudogenes [19]. Other six genes are known for their pseudogene mediated expression regulation or for their involvement in cancer disease.

The proposed algorithm was able to detect 110 of 121 pseudogenes annotated by Ensembl for these 11 genes. Moreover, it detected four loci not reported by Ensembl, but reported by UCSC, two new potential pseudogene loci reported neither by Ensembl nor by UCSC and three duplicated sequences for three distinct pseudogenes. Though the algorithm did not capture all the annotated pseudogenes, it seems to be an efficacious solution to detect new potential loci, especially when the query coverage of the alignment is shorter than the coding sequence. These loci can be classified to pseudogene fragments.

The downstream regions of the detected pseudogenes were analyzed in order to find polyA tails. We found a polyA tail for 48 pseudogenes and a polyadenylation signal for 13 of them. These numbers are coherent with known data. Literature reports that a polyA tail is present in about 45-50% of the cases [19].

CpG islands of different lengths and at different distances from the pseudogenes were detected in 16 upstream regions. We didn't find

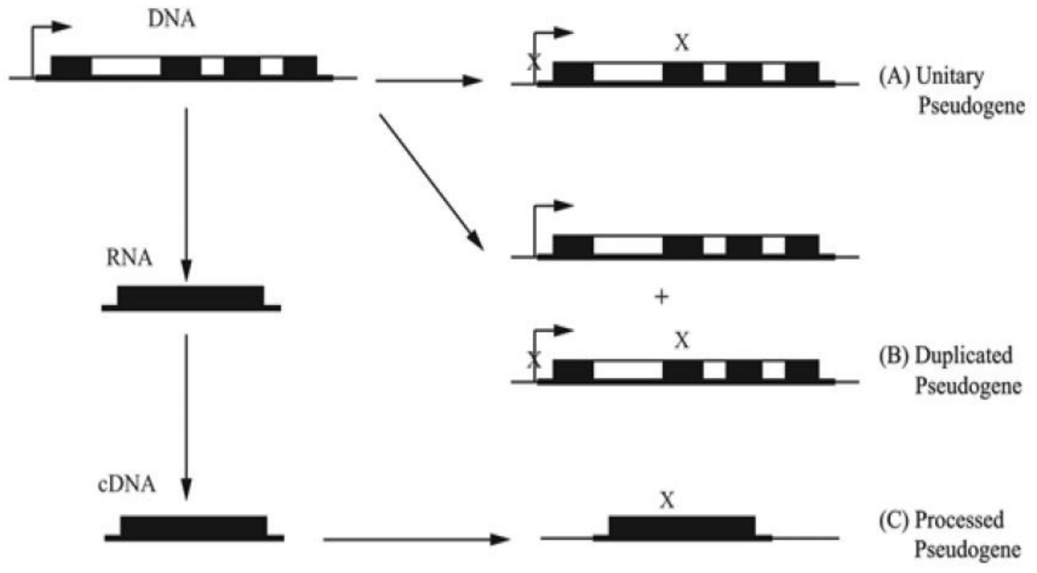

Figure 1: (A) Unitary pseudogenes, (B) Duplicated pseudogenes, (C) Processed pseudogenes [29].

Citation: Trucco G, Cerioli V (2021) Detection and Analysis of Pseudogenes in Non-coding DNA. Int J Comput Softw Eng 6: 164. doi: https://doi.org/10.15344/2456-4451/2021/164

any motif in the upstream regions probably because a bigger set of sequences is needed by the Gibbs sampling algorithm. However, we executed the algorithm on the flanking regions of some pseudogenes and the results showed an interesting similarity between the flanking regions of some of them. These similarities were confirmed also by alignments of the regions.

Conclusion and prospects

The complex and not yet elucidated system of interrelations between parental gene mRNAs and pseudogenes transcripts could give reasons for the complexity of vertebrates beyond their genome sizes [29]. Moreover, the comprehension of these mechanisms could bring to light new treatments for many diseases, from cancer to diabetes.

Detection of pseudogenes based on raw nucleotide identity should be applied also for the duplicated pseudogenes. These are longer than processed pseudogenes because they include also introns but, unlike processed pseudogenes, they usually lie in the vicinity of their parental genes and, therefore, their detection does not need the scanning of the entire genome [12].

It was observed that CpG islands not associated with a known promoter (orphan CpG islands) are active transcription initiators during development and, after that, they lose this feature. This evidence suggests that orphan CpG islands are associated with undetected promoters [43]. The search for unknown promoter sequences near the orphan CpG islands in the pseudogenes upstream regions is a promising future development of this work. Moreover, the similarity between the flanking regions of some pseudogenes raises questions about their origin.

Methods

In this section, we provide detailed descriptions of the algorithms and the strategies used in this project to pursue the following goals:

1. identification of the processed pseudogenes of some selected genes;
2. detection of a polyA tail in the downstream region of each identified pseudogene;
3. detection of CpG islands in the upstream region of each pseudogene;
4. motif discovery (search for potential promoters sequences) in the upstream regions of the pseudogenes.

Identification

The first step is design of a strategy for identifying the pseudogenes of a gene starting from its CDS. We developed a program that scans the entire genome and stores all the sequences that have similarity with a selected CDS in terms of transition probabilities (the probabilities of transition between the different nucleotides in the CDS) and occurrences of the nucleotides. Each stored sequence is then aligned with the CDS and, if the alignment is statistically significant, the sequence is marked as a pseudogene.

As a first step, the program builds a matrix of the transition probabilities of the CDS and computes the probability of the CDS itself according to this model (CDS_P). The probability is computed as a sum of logarithms of probabilities in order to avoid floating point underflow errors (that is numbers of smaller absolute values than the computer can represent in its CPU) or, worse, the production of arbitrary wrong numbers. The nucleotides occurrences of the CDS (CDS_{C_n}) are also calculated. A sliding window that has the same length of the CDS scans the entire genome. When it finds a sequence with a transition probability that is included in the interval $CDS_P \pm CDS_P \cdot 0.05$ and a nucleotide occurrence in the interval $CDS_{C_n} \pm CDS_{C_n} \cdot 0.2$, for each nucleotide, the ends of the sequence are stored in a list. The window can enlarge itself until the above-mentioned conditions are satisfied. Sequences longer than four times the CDS length will be discarded from the list at the end of the scanning. A distinct program builds 100.000 random sequences with the same transition probabilities of the CDS. Each sequence is aligned with the CDS and the program returns the mean and the standard deviation of the alignment scores. Then we align the CDS with all the sequences in the list. For each alignment, the main program computes the z-score given by

$$Z = \frac{X - \mu}{\sigma},$$

using the mean μ and the standard deviation σ previously computed as explained above. A threshold of 8 is chosen for the z-score so that only sequences with a z-score greater than the threshold are recorded as pseudogenes. We chose a threshold of 8 because the alignment scores between the CDSs and the random sequences are not normally distributed [34]. The parameters of the alignment algorithm are: match = 1, mismatch = 0 and gap = -1. Figure 2 summarizes the identification strategy described in this paragraph.

PolyA tails

A polyA tail is a stretch of RNA that has only adenine bases. In eukaryotes, the addition of a polyA tail to a messenger RNA 3' end is part of a process that produces mature messenger RNA (mRNA) and is called polyadenylation. Processed pseudogenes are typically characterized by the lack of introns and the presence of residue of the polyA tail, unless it has not decayed [19]. We searched for a polyadenine tail by means of a 50 bp sliding window in the 1000 bp (base pairs) length region beyond the pseudogene 3' end. The 50 bp windows containing more than 30 adenines are memorized (if they exist) and the most promising one is considered as a PolyA tail. When a polyadenine tail is found, the algorithm searches for a polyadenylation signal (AATAAA or ATTAAA) in the 100 bp length upstream region of the tail.

CpG islands

CpG islands are regions of DNA in which a cytosine is followed by a guanine in the linear sequence of nucleotides along the $5' \rightarrow 3'$ direction with a high frequency [48]. The notation CpG is used to distinguish the single strand sequence from the CG pairing on the double strand [49]. In vertebrate genomes, CpG nucleotides occur with a much lower frequency than would be expected by random chance. The frequency of CpG dinucleotides in the human genome is 0.98% while the expected frequency is 4.41%. CpGs cytosines are often methylated (70%). The very low occurrence of CpGs is explained by the fact that methylated cytosines are mutational hotspots and this has led CpGs depletion during evolution [45]. CpG islands usually occur near the transcription start site of genes and have an important role in gene expression regulation. While methylated CpG islands inhibit transcription, unmethylated CpG islands near a transcription start site enables the transcription of that gene. Consequently, CpG islands play an important role in gene expression regulation and the ability

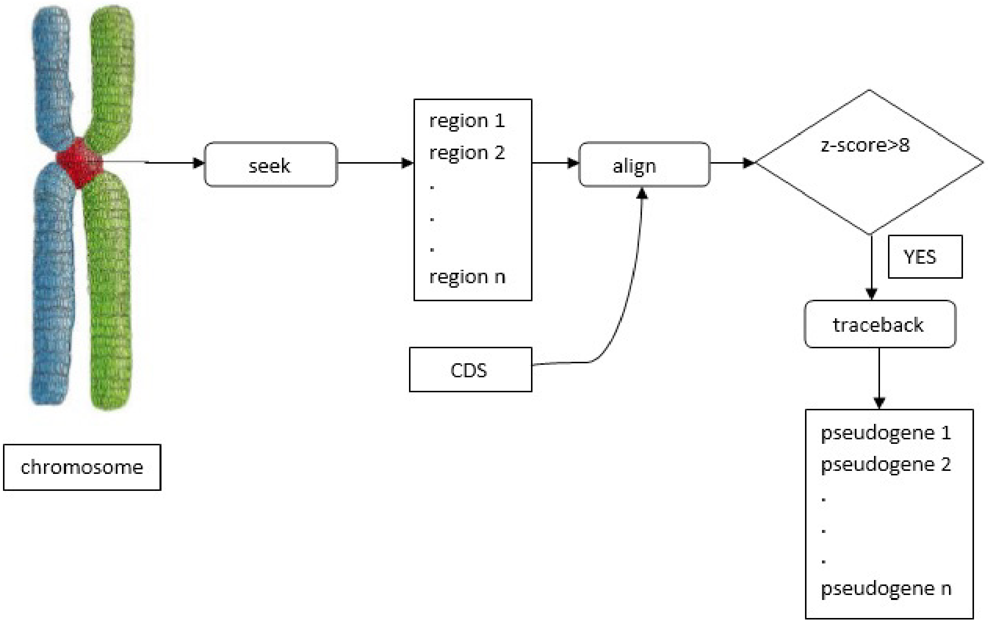

Figure 2: Each chromosome is scanned and the coordinates of all the regions that are homologous to the CDS are stored in the list. Each homologous region is then aligned with the CDS and, if the z-score of the alignment is greater than 8, the region is recorded as a potential pseudogene.

to identify them can help us to predict the location of genes within the DNA. A naïve approach to locate CpG islands in a sequence X of length L is to extract a sliding window of length $len \ll L$ and to compute a score for each subsequence of length *len* in X. The main disadvantage of this strategy is that we have no information about the length of the islands. If we use a value of *len* that is too large, the score we get from this window may not be high enough. The best approach for this problem is the use of a HiddenMarkov Model (HMM). A general HMM [35] is a triplet

$$M = (Q, S, \Theta),$$

where:

1. Q is an alphabet of symbols;
2. S is a finite set of states capable of emitting symbols from the alphabet Q;
3. Θ is a set of probabilities, comprised of:

1. state transition probabilities, denoted as p_{ij} for each $i, j \in S$;
2. emission probabilities denoted as $q_k(b)$ for each $k \in S$ and $b \in Q$.

The HMM for CpG islands has [36]:

1. 9 states: begin/end, A+, C+, G+, T+, A-, C-, G and T-
2. 4 symbols: A, C, G and T

The letters A+, C+, G+ and T+ represent states that belong to a CpG island. The other letters, instead, represent states not belonging to a CpG island. The state 0 corresponds to the state begin/end of the chain. A Markov chain is a system (S, A) consisting of a finite set of states S and a transition matrix $A= a_{kl}$ with $\sum_{l \in S} a_{kl} = 1$ for all $k \in s$ that determines the probability of the transition $k \rightarrow l$ by $P(s_{i+1} = l \mid s_i = k) = a_{kl}$. At any step i, the Markov chain is in a specific state s_i and the chain changes to state s_{i+1} according to the given transition probability [36].

	A	C	G	T
0	0	0	0	0
A+	1	0	0	0
C+	0	1	0	0
G+	0	0	1	0
T+	0	0	0	1
A-	1	0	0	0
C-	0	1	0	0
G-	0	0	1	0
T-	0	0	0	1

Table 1: HMM for CpG islands.

Table 1 reports the emission probability matrix. It's worth to notice that, in this model, each state emits only the corresponding symbol/nucleotide (with probability 1).

The state transition probabilities matrix is reported in Table 2. Model "+" describes the transition probabilities inside the CpG, model "-" describes the transition probabilities outside the CpG island [36].

The Viterbi algorithm takes as input the query sequence X, the transition probabilities, the emission probabilities of the model and returns a path π that maximizes $P(X, \pi)$ (may not be unique) [37]. This value is the probability of the most probable path, that is the underlying chain of the HMM. This search performed with HMMs is called *decoding* and the paths found are called *Viterbi paths*. Suppose we have a HMM with a finite state $S = \{G_1,\ldots,G_N\}$ and transition probabilities $P_{0i}, P_{ij}, P_{j0} \forall i, j = 1, \ldots, N.$ Given a sequence of letters $X = X_1 \ldots X_L$ from the alphabet Q, we define

$$v_k(1) = p_{0k} q_k(X_1)$$

for $k = 1,...,N$. Then we have: $v_k(i) = \max_{\pi_1 ... \pi_{i-1}} \in X\ p_{0\pi_1} q_{\pi_1}(x_1) p_{\pi_1\pi_2} q_{\pi_2}(x_2) \times ... \times p_{\pi_{i-1}} G_k q_k(x_i)$, for $i = 2,...,L$ and $k = 1,...,N$.

Now we can compute the probability of each step using the initial condition $v_k(1)$ and the following recursive step:

$$v_k(i+1) = q_k(X_{i+1}) \max_{l=1,...,N}(v_l(i) p_{lk}).$$

At the end of the execution, we have the probability of the most probable path given by

$$\max_{l=1,...,N}(v_l(L) p_{l0}).$$

Since in our model each letter can be emitted by only two states (i+ and i-), we have a simplified version of the algorithm in which, for $i = 2,..L$, $p_{\pi_{i-1}\pi_i} = p_{\pi_{i-1}\pi_{i+}}$ or $p_{\pi_{i-1}\pi_i} = p_{\pi_{i-1}\pi_{i-}}$ and $q_k(X_i) = 1$. To reconstruct the sequence of states, that is the Viterbi path, at each step we form a set $V_k(i)$ made of all integers m for which $v_m(i)p_{mk} = \max_{l=1,...,N}(v_l(i)p_{lk})$. The Viterbi path can be recovered by recursively choosing $m_L \in V\ (L)$, then $m_{L-1} \in V_{mL}\ (L-1)$ *until* $m_1 \in V_{m2}\ (1)$.

Motif discovery

In order to find potential sequence signals (DNA binding sites or promoters) in the upstream regions in which a CpG island is present, we developed a Gibbs sampling algorithm capable of locating a pattern of subsequences with the highest likelihood. Gibbs sampling is a probabilistic inference algorithm used to generate a sequence of samples from a joint probability distribution of two or more random variables [38]. In bioinformatics, Gibbs sampling is used to detect motif signals in multiple DNA or protein sequences assuming no prior information about the motifs [32,33,39]. Thus, given a set of sequences $S=S^{(1)},..,S^{(n)}$ and an integer w, the algorithm finds, for each sequence $S^{(i)}$, a subsequence of length w, so that the similarity between the n sequences is maximized [39,40]. Let c_{ij} be the number of occurrences of the symbol $j \in \Sigma$ among the i^{th} position of the n subsequences. Let q_{ij} denote the probability of the symbol j to occur at the i^{th} positions of pattern and let p_j denote the frequency of the symbol j in all sequences of S. The algorithm maximizes the equation:

$$F = \sum_{i=1}^{w} \sum_{j \in \Sigma} c_{ij} . \log \frac{q_{ij}}{p_j},$$

where c_{ij} and q_{ij} are computed from the complete alignment of the subsequences. To achieve this result, we designed an algorithm that performs the following iterative procedures:

1. Initialization: randomly chooses $a^{(1)},...,a^{(n)}$, the starting indices of the subsequences in $S^{(1)},...,S^{(n)}$, respectively.
2. Randomly chooses $1 \leq z \leq n$ and computes c_{ij}, q_{ij} and p_j values for the sequences in $S \backslash S^{(z)}$.
3. According to the model, computes the weights of all possible subsequences of length w in $S^{(z)}$. The weights are normalized and a new value of $a^{(z)}$ is randomly selected with a probability proportional to the weights of the subsequences of $S^{(z)}$. In order to avoid local optima, the starting position with the highest weight is not guaranteed to be chosen. In order to rapidly converge to a solution, the above mentioned random sampling goes on for a fixed amount of time (usually 15 min), then, after the time threshold has expired, only the position with the highest weight is chosen.
4. The algorithm repeats step 2 and 3 until it converges to a fixed pattern of subsequences. The algorithm ends when the same pattern of subsequences is produced for 10 consecutive iterations.

We chose this strategy with the purpose of having many "fast" solutions rather than few "slow" ones.

Results and Discussion

We implemented the algorithms in Java language and we executed them on a Notebook Asus K72F equipped with Intel Core i3 processor (2.5 GHz). The entire human genome sequence was downloaded from the repository on www.ncbi.nlm.nih.gov, the CDSs were downloaded from the Ensembl genome browser hosted by www.ensembl.org. In this section, we describe the results of the following experiments.

1. In order to detect the pseudogenes of each gene considered in the survey, we developed a strategy based on raw nucleotide identity that scans the entire human genome and returns the coordinates of each detected pseudogene.
2. The 1000 bp length downstream region of each pseudogene was inquired about the presence of a polyA tail and, when this feature was present, the algorithm searched for a polyadenylation signal in the 100 bp length upstream region of the tail.
3. The 1000 bp length upstream region of each pseudogene was decoded by the Viterbi algorithm based on a HMM suited for CpG islands detection.
4. We also performed motif discovery experiments on the flanking regions of some pseudogenes. The strategy used for this goal was Gibbs sampling.

In our research we identified and analyzed the pseudogenes of the following genes:

1. RPL14, RPL19, RPL22, RPL36 and RPL37 are ribosomal protein genes (RP family). It was discovered that the protein family that has the largest number of processed pseudogenes is the RP family (more than 2000) [12].

	0	A+	C+	G+	T+	A-	C-	G-	T-
0	0.000	0.0725193	0.1637630	0.1788242	0.0754545	0.1322050	0.1267006	0.1226380	0.1278950
A+	0.001	0.1762237	0.2682517	0.4170629	0.1174825	0.0035964	0.0054745	0.0085104	0.0023976
C+	0.001	0.1672435	0.3599201	0.2679840	0.1838722	0.0034131	0.0073453	0.0054690	0.0037524
G+	0.001	0.1576223	0.3318881	0.3671328	0.1223776	0.0032167	0.0067732	0.0074915	0.0024975
T+	0.001	0.0773426	0.3475514	0.3750440	0.1781818	0.0015784	0.0070929	0.0076723	0.0036363
A-	0.001	0.0002997	0.0002047	0.9992837	0.0002097	0.2994005	0.2045904	0.2844305	0.2095804
C-	0.001	0.0003216	0.0002977	0.0000769	0.0003016	0.3213566	0.2974045	0.0778441	0.3013966
G-	0.001	0.0001768	0.0002387	0.0002917	0.0002917	0.1766463	0.2385224	0.2914165	0.2914155
T-	0.001	0.0002477	0.0002457	0.0002977	0.0002077	0.2475044	0.2455084	0.2974035	0.2075844

Table 2: Transition matrix.

2. PTEN (phosphatase and tensin homolog) codes for a tumor suppressor. PTENP1, the PTEN pseudogene, is transcribed and shares close homology with the mature PTEN mRNA and, as a consequence, it can act as a sponge for some specific micro RNAs (miRNA) removing their repression of PTEN expression. Since it was observed that little changes in PTEN protein levels can lead to tumor, PTENP1 can be considered a tumor suppressor gene in its own right. In melanoma, in fact, the PTENP1 is suppressed [25].
3. KRAS (GTPase KRAS) is a proto-oncogene. Overexpression of KRASP1, the KRASP transcribed pseudogene, causes an increased level of KRAS mRNA and this accelerates cell growth. It was discovered that KRAS and KRASP1 transcript levels are positively correlated in prostate cancer [23].
4. RAP1A and RAP1B are members of the oncogene RAS family.
5. CX43 (gap junction protein alpha) is another cancer-related gene. GJA1P1, a pseudogene of CX43, is expressed in breast cancer but not in normal cells [41].
6. HDAC1 (histone deacetylase 1) expression is regulated by the pairing of two transcribed pseudogenes, one transcribed in the sense direction, the other in the antisense direction. The double stranded RNA produced by this pairing leads to the degradation of the mRNA from the parental coding gene [42].

Pseudogenes detection

The Ensembl genome browser reports 121 pseudogenes for these 11 genes. We attested 110 of them and we identified 6 pseudogenes loci not previously annotated by the Ensembl genome browser, two of them annotated neither by the Ensembl genome browser nor by the UCSC genome browser (www.genome.ucsc.edu). The statistical significance of the alignments was confirmed by the z-score and by the BLASTN alignment online application hosted by the National Center for Biotechnology Information (NCBI) website (www.ncbi.nlm.nih.gov). The position and the annotation of the sequences found were confirmed by the Ensembl genome browser. Table 3 reports, for each gene, the number of pseudogenes annotated by Ensembl (first row), the number of loci attested by our method (second row) and the loci not reported by Ensembl (third row), but detected by our method.

Tables 4, 5, 6 and 7 show the detection results for RPL14, RPL19, RPL22 and RPL37. The first column reports the position in the genome sequence (number of the chromosome, where +1 stands for forward strand and -1 for reverse strand) of the gene itself and of each detected pseudogene. The second column reports the z-score of the alignments. The third and the fourth columns report the query coverage and the percent identity of the alignments respectively. The latter two parameters are provided by BLASTN.

The computation time of pseudogenes detection depends on CDS length because the optimal alignment is computed in *O(L^2)*, where *L* is the length of the sequence. However, the main factor that influences the computation time is the number of homologous sequences found, which is unknown before execution. Table 8 reports the computation time of each experiment.

PolyA tails

We found 48 polyA tails (41% of the cases) and 13 polyadenylation signals (AATAAA or ATTAAA). It's worth to notice that the sequence (1) of RPL37, reported neither by Ensembl nor by UCSC, has a polyA tail at 571 bp from its 3' and a polyadenylation signal at 13 bp from the 5' of the tail. Table 9 shows the number of tails and the number of polyadenylation signals found for each group of pseudogenes.

	RPL14	RPL19	RPL22	RPL36	RPL37	PTEN	KRAS	RAP1A	RAP1B	CX43	HDAC1	sum
annotated	9	22	23	25	28	2	1	2	5	1	3	121
attested	9	19	22	21	26	1	1	2	5	1	3	110
not reported	1	2	2	0	1	0	0	0	0	0	0	6

Table 3: The first row reports the number of annotated pseudogenes for each gene, the second and the third rows report the number of attested pseudogenes and of unannotated pseudogene loci identified by our method respectively.

	chromosome	z-score	query coverage	percent identity
RPL14	3 (+1)			
RPL14P1	12 (+1)	69.86	84%	98.21%
AC017079.1	2 (+1)	24.92	42%	86.44%
AC012519.1	3 (-1)	42.78	84%	83.89%
AC126615.1	12 (+1)	56.17	82%	91.28%
RPL14P3	4 (-1)	53.20	83%	90.50%
AC108039.1	2 (-1)	17.78	37%	90.00%
AL024507.1	6 (-1)	13.32	43%	82.95%
RPL14P5	X (-1)	34.15	58%	85.58%
AC117522.3	5 (-1)	26.41	45%	86.96%
AL356967.1	6 (+1)	9.75	30%	71.53%

Table 4: The locus AL356967.1 is annotated by Ensembl as "novel pseudogene" residing on chromosome 6 (forward strand) at 104.687.241-104.687.879. UCSC reports it as RPL14 retrogene.

	chromosome	z-score	query coverage	percent identity
RPL19	17 (+1)			
AL161742.1	1 (-1)	88.28	99%	91.36%
AC097358.1	3 (+1)	86.98	99%	91.31%
AL359092.2	9 (+1)	83.09	99%	90.44%
AC133435.1	3 (-1)	84.07	99%	91.54%
AC092824.1	12 (-1)			
AC021660.1	3 (+1)			
AC095041.1	4 (+1)	86.33	100%	90.37%
AC010326.1	19 (-1)	72.08	98%	83.32%
AC136632.1	5 (+1)	99.94	99%	97.29%
AC007683.1	7 (-1)	95.73	99%	95.25%
AC068137.1	2 (+1)	85.36	98%	90.91%
AC091980.1	5 (-1)	88.28	99%	91.50%
AC009997.1	15 (+1)	14.11	23%	93.43%
AC021016.1	2 (+1)	74.03	85%	91.88%
RPL19P20	X (+1)	87.95	99%	91.17%
RPL19P21	X (+1)	94.11	99%	94.24%
RPL19P13	8 (+1)	81.80	99%	88.44%
RPL19P14	8 (+1)	79.21	97%	87.35%
RPL19P18	17 (-1)	71.11	99%	84.09%
RPL19P11	5 (+1)	86.66	99%	90.36%
RPL19P16	10 (-1)	76.29	99%	89.59%
RPL19P1	20 (+1)			
(1)	3 (+1)	23.82	60%	73.90%
(2)	10 (+1)	88.93	99%	91.86%

Table 5: The sequence (1) coincides to an intronic sequence of PAK2 (p21 activated kinase 2) on chromosome 3 (forward strand) at 196.823.591-196.824.193. The sequence (2) is aligned by the Ensembl gene browser with the 3' portion of the novel transcript FP565171.1 on chromosome 10 (forward strand) at 131.895.114-131.895.702. They are both annotated by UCSC as RPL19 retrogenes, but not by Ensembl. Two duplicated sequences, one of AC007683, the other of RPL19P13, were found on chromosomes 7 (forward strand) and 8 (reverse strand) respectively (not shown in the table). Three pseudogenes of RPL19 were not detected (AC092824.1, AC021660.1 and RPL19P1).

	chromosome	z-score	query coverage	percent identity
RPL22	1 (-1)			
RPL22P19	12 (-1)	61.84	99%	90.98%
RPL22P11	2 (-1)	64.98	99%	93.01%
AC012443.1	2 (-1)	26.93	52%	82.38%
RPL22P18	10 (+1)	56.97	96%	81.63%
RPL22P24	1 (-1)	50.34	99%	85.05%
RPL22P23	X (-1)	49.30	90%	87.75%
RPL22P7	2 (-1)	51.39	99%	86.53%
RPL22P17	10 (-1)			
RPL22P20	14 (+1)	58.71	99%	90.41%
RPL22P1	3 (-1)	62.89	91%	98.03%
RPL22P10	2 (+1)	59.74	95%	91.91%
RPL22P22	X (-1)	58.71	99%	90.16%
RPL22P16	7 (-1)	59.41	99%	91.45%
RPL22P12	2 (+1)	64.28	99%	92.51%
RPL22P4	1 (-1)	56.62	99%	88.07%
RPL22P3	1 (-1)	61.84	98%	91.58%
RPL22P5	1 (-1)	56.62	99%	88.07%
RPL22P6	1 (-1)	56.97	99%	88.32%
RPL22P8	2 (+1)	63.59	99%	91.97%
RPL22P21	17 (+1)	38.00	71%	83.45%
RPL22P2	14 (-1)	63.94	95%	94.88%
RPL22P13	4 (+1)	15.84	66%	71.12%
RPL22P14	6 (+1)	43.02	98%	82.12%
(1)	11 (-1)	31.17	86%	90.98%
(2)	21(-1)	16.18	40%	79.38

Table 6: The sequence (1) is located in an intronic sequence of PHF21A (PHD finger protein 21A) on chromosome 11 (reverse strand) at 46.053.694-46.054.066 and it is annotated by UCSC as RPL22 retrogene, but not by Ensembl. The sequence (2) is located on chromosome 21 (reverse strand) at 41.135.115-41.135.527. This sequence is annotated neither by UCSC nor by Ensembl. We can consider this sequence as a fragment because of its low query coverage in the BLASTN alignment (40%). The algorithm found one duplicated sequence of RPL22P14 on the forward strand of chromosome 6 (not shown). The pseudogene RPL22P17 was not detected.

	chromosome	z-score	query coverage	percent identity
RPL37	5 (-1)			
AC010655.1	7 (-1)	52.24	99%	87.71%
AL451007.1	1 (-1)	58.54	99%	93.49%
AC006512.1	12 (+1)	51.40	98%	95.90%
AC098869.1	4 (+1)	51.40	98%	86.94%
AC004223.1	17 (-1)	41.33	88%	93.02%
AP001922.4	11 (+1)	42.59	91%	83.81%
AL359092.1	9 (-1)	35.03	89%	88.27%
AL589872.1	X (+1)	45.94	99%	82.19%
AC015977.1	2 (-1)	56.02	99%	90.78%
AL590128.1	1 (-1)	49.30	93%	87.59%
AC091133.6	17 (-1)	37.97	69%	91.13%
AL355598.2	10 (+1)	23.70	62%	85.52%
AL049874.2	14 (-1)	52.66	93%	89.82%
AC011753.4	2 (+1)			
RPL37P18	10 (-1)	19.92	54%	81.76%
RPL37P21	13 (-1)			
RPL37P23	19 (+1)	63.15	99%	95.90%
RPL37P10	2 (+1)	43.43	99%	80.89%
RPL37P6	8 (+1)	61.89	99%	94.88%
AL049745.2	1 (+1)	50.14	98%	86.55%
AC004801.2	12 (-1)	51.40	98%	87.29%
RPL37P4	21 (+1)	47.20	99%	84.98%
RPL37P2	11 (-1)	63.15	99%	95.90%
RPL37P25	5 (+1)	28.73	72%	80.84%
RPL37P3	21 (-1)	51.40	99%	86.69%
RPL37P15	6 (-1)	54.76	97%	89.93%
RPL37P1	20 (+1)	50.14	99%	88.74%
AC107083.1	2 (-1)	40.07	99%	80.55%
(1)	11 (+1)	19.92	45%	89.63%

Table 7: The sequence (1) is located on chromosome 11 (forward strand) at 75.644.799-75.645.091 and it is annotated neither by UCSC nor by Ensembl. Given its low query coverage (45%) and high percent identity (89.63%) in the BLASTN alignment, this locus can be regarded as a pseudogene fragment. The algorithm did not detect AC011753.4 and RPL37P21.

gene	RPL14	RPL19	RPL22	RPL36	RPL37	PTEN	KRAS	RAP1A	RAP1B	CX43	HDAC1
CDS length	552	591	387	320	294	1212	570	555	555	1146	1449
time (min)	86	70	124	46	102	422	179	254	342	192	273

Table 8: The table shows the length of each CDS and the computation time needed to scan the entire genome in search of its pseudogenes.

	RPL14	RPL19	RPL22	RPL36	RPL37	PTEN	KRAS	RAP1A	RAP1B	CX43	HDAC1	sum
analyzed	10	21	24	21	27	1	1	2	5	1	3	116
tail	4	11	8	12	12	0	0	1	0	0	0	48
signal	1	3	4	1	4	0	0	0	0	0	0	13

Table 9: The first row reports the number of pseudogenes analyzed for each gene (attested+not reported), the second and the third rows report the number of tails and the number of polyadenylation signals for each group respectively.

CpG islands

The upstream 1000 bp length regions of the detected pseudogenes were analyzed in order to check the presence of CpG islands. Table 10 displays the number of CpG islands found for each group of pseudogenes and the maximum CpG island length in each group. It is worth to notice that the length of the CpG islands varies a lot and, therefore, a sliding window cannot be used to detect CpG islands.

Motif discovery and flanking regions

It was observed that half of mammalian CpG islands (~ 10.000) are "orphan", that is, they are not associated with annotated promoters. There are evidences that many orphan CpG islands play a role as transcriptional initiator during development and, after that, they are subject to DNA methylation loosing their active promoter features. Thus, orphan CpG islands may correspond to undetected promoters that are active during development [43]. With the aim of finding a possible DNA signal in the CpG islands found, we analyzed the 500 bp length pseudogenes upstream regions that contain CpG islands. We run the Gibbs sampling algorithm in order to find common subsequences of length 14. We didn't find any significant common motif. Nevertheless, we identified a similarity between the upstream regions of RAP1B pseudogenes. We noticed that the subsequences of the best pattern for these regions are located at similar distances from their respective pseudogenes 5' ends. The same happens for the subsequences of other high-scored patterns. A similar feature was observed also in the downstream regions (excepting AL161670.1). This feature was not observed in the upstream (and downstream) sequences of the pseudogenes of RAP1A, PTEN and HDAC1. We did not test the ribosomal pseudogenes for this feature. Table 11 shows the distances from the 5' ends of the RAP1B pseudogenes of the three best upstream patterns. Table 12 shows the distances from the 3' ends of the RAP1B pseudogenes of the three best downstream patterns.

A further confirmation of the similarity between the flanking regions of these pseudogenes is provided by the alignment BLASTN online application hosted by the NCBI website. In Table 13, the bottom-left triangle contains the alignments scores (qc = query coverage and pi = percent identity) of the upstream regions. The top-right triangle contains the results of the downstream regions alignments.

	first	second	third
AC113404.3	142	123	64
RAP1BP1	127	107	55
RAP1BP2	121	104	55
RAP1BP3	134	114	55
AL161670.1	137	117	55

Table 11: The table shows the distances of the three bestscored upstream patterns from the 5' ends of the pseudogenes of RAP1B.

	first	second	third
AC113404.3	267	25	78
RAP1BP1	266	25	78
RAP1BP2	265	25	74
RAP1BP3	268	24	79
AL161670.1	195	138	365

Table 12: The table shows the distances of the three bestscored downstream patterns from the 3' ends of the pseudogenes of RAP1B.

Conclusion

Though the genomes of higher organisms do not have more genes than lower organisms, the greater abundance of regulatory ncRNAs, found in the higher organisms, could give reasons to a more complex phenotype from the same building blocks [29]. Characterizing the pseudogenes and understanding their regulatory role will help in discovering the genetic background of many diseases but also in identifying new pharmacological treatments. Moreover, the prevalence of pseudogenes in mammalian genomes can introduce artifacts in automatic gene annotation pepelines in which pseudogenes are often mistakenly annotated as genes. This is due to the high sequence similarity of pseudogenes with their parental genes [12, 44]. Therefore, the correct identification of pseudogenes is important also for gene annotation.

Identification: No consensus computational scheme for detecting and defining pseudogenes has yet been developed. Distinct pseudogene annotation strategies produced rather distinct set of pseudogenes [12]. The algorithm based on raw nucleotide identity, even if it did

	RPL14	RPL19	RPL22	RPL36	RPL37	PTEN	KRAS	RAP1A	RAP1B	CX43	HDAC1	sum
analyzed	10	21	24	21	27	1	1	2	5	1	3	116
CpG	0	2	2	4	1	1	1	2	3	0	0	16
max len.		132	142	268	90	805	95	100	111			

Table 10: The first row reports the number of pseudogenes analyzed for each gene (attested+not reported), the second row reports the number of CpG islands identified in each group. The last row displays the maximum CpG island length in each group.

	AC113404.3	RAP1BP1	RAP1BP2	RAP1BP3	AL161670.1
AC113404.3		qc=100%, pi=92.83%	qc=100%, pi=87.23%	qc=99%, pi=91.40%	no significant similarity
RAP1BP1	qc=32%, pi=83.45%		qc=100%, pi=82.47%	qc=99%, pi=86.17%	no significant similarity
RAP1BP2	qc=16%, pi=73.33%	qc=31%, pi=70.34%		qc=100%, pi=81.27%	qc=2%, pi=100%
RAP1BP3	qc=35%, pi=72.15%	qc=27%, pi=70.75%	qc=28%, pi=65.94%		no significant similarity
AL161670.1	qc=32%, pi=89.44%	qc=33%, pi=82.68%	qc=18%, pi=85.07%	qc=2%, pi=92.26%	

Table 13: The table reports the scores of the alignments among the upstream regions (bottom-left) and among the downstream regions (top-right) of the pseudogenes of RAP1B.

not "capture" all the pseudogenes annoted by Ensembl, proved to be an efficacious tool for detection of new potential pseudogene sites not discovered by other strategies. In particular, it seems capable to cut off statistically significant alignments with a low query coverage, which we can regard as pseudogene fragments [19]. The algorithm parameters (thresholds for the transition probability and for the nucleotides occurrences), which we chose empirically, have to be refined in order to improve the performance of the algorithm. Moreover, it should be tested also for detection of duplicated pseudogenes. These are longer than processed pseudogenes because they include introns. However, unlike processed pseudogenes, they reside near their parental genes and, as a consequence, they do not need the scanning of the entire genome to be detected [12].

PolyA tails: Although it was observed that a polyA tail is present beyond a processed pseudogene in about half of the cases [19], the presence of a polyA tail (with a possible polyadenylation signal) could help the definition of a sequence as a processed pseudogene.

CpG islands and motif discovery: The accepted definition of what is a CpG island was proposed in 1987 as being a 200 bp stretch of DNA with a C+G content of 50% and an observed CpG/expected CpG in excess of 0.6 [45]. However, any definition of CpG island, after all, is arbitrary [46]. Using a HMM designed for the purpose, we found some CpG islands of different lengths and located at different distances from the pseudogenes. Then we tried to find a motif (or signal) in the upstream regions in which a CpG island is present. The issue of searching for possible promoter sequences within these orphan CpG regions is a promising future development of this work. The experiments with the Gibbs sampling showed a surprising similarity between the flanking regions of some pseudogenes of the same gene. This suggests that generation of the processed pseudogenes should be further investigated.

Competing Interests

The authors declare that they have no competing interests.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome. Nature 409: 860-921.
2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic localization of common disease associated variation in regulatory DNA. Science 337: 1190-1195.
3. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M, et al. (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22: 1748-2759.
4. Martinez AF, Abe Y, Hong S, Molyneux K, Yarnell D, et al. (2016) An ultraconserved brain-specific enhancer within DGRL3 (LPHN3) underpins attention-deficit/hyperactivity disorder susceptibility. Biol Psychiatry 80: 943-954.
5. Amiel J, Benko S, Gordon CT, Lyonnet S (2010) Disruption of long-distance higly conserved noncoding elements in neurocristopathies Ann N Y Acad Sci 1214: 34-46.
6. Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, et al. (2011) Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. Proc Natl Acad Sci USA 108: 786-791.
7. Bao BY, Lin VC, Yu CC, Yin HL, Chang TY, et al. (2016) Genetic variants in ultraconserved regions associate with prostate cancer recurrence and survival. Scientific Reports 6: 22124.
8. Zhang Fand, Lupsky JR (2015) Non-coding genetic variants in human disease. Hum Mol Genet 24: R102-R110.
9. Eddy SR (2014) The C-value paradox, junk DNA and ENCODE. Curr Biol 22: R898-R899.
10. Palazzo AF, Gregory TR (2014) The case for junk DNA. PLoS Genet 10: e1004351.
11. Koonin EV (2005) Orthologs, Paralogs and Evolutionary Genomics. Annu Rev Genet 39: 309-338.
12. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, et al. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17: 839-851.
13. Niimura Y, Nei M (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2: 860-921.
14. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, et al. (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? RNA 17: 792-798.
15. Korneev SA, Park JH, O'Shea M (1999) Neuronal Expression of Neural Nitrix Oxide Synthase (nNOS) Protein is suppressed by an Antisense RNA Transcribed from an NOS Pseudogene. J Neurosci 19: 7711-7720.
16. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M, et al. (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 11: R26.
17. Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS letters 468: 109-114.
18. D'Errico I, Gadaleta G, Saccone C (2014) Pseudogenes in metazoa: origin and features. Brief Funct Genomic Proteomic 3: 157-167.
19. Zhang Z, Harrison P, Gerstein M (2002) Identification and analysis of over 200 ribosomal protein pseudogenes in the human genome. Genome Res 12: 1466-1482.
20. McDonell L, Drouin G (2012) The abundance of processed pseudogenes derived from glycolitic genes is correlated with their expression level. Genome 5: 147-151.
21. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103: 3220-3225.
22. Lin M, Pedrosa E, Shah A, Hrabovsky A, Maqbool S, et al. (2011) RNA-seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders. PLoS Oone 6: e22356.
23. Poliseno L, Salamena L, Zhang J, Carver B, Haveman WJ, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumor biology. Nature 465: 1033-1038.
24. Zhou M, Baitei EY, Alzahrani AS, Al-Mohanna F, Farid NR, et al. (2009) Oncogenic activation of MAP kinase by BRAF pseudogene in thyroid tumors. Neoplasia 11: 57-65.
25. Chiefari E, Iiritano S, Paonessa F, Pra IL, Arcidiacono B, et al. (2010) Pseudogene-mediated pstrascriptional silencing of HMGA1 can result in insulin resistance and Type 2 diabetes. Nat Commun 1: 40.
26. Khachane AN, Harrison PM (2009) Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics 10: 435.
27. Podlaha O, Zhang J (2004) Nonneutral evolution of the transcribed pseudogene Makorin1-p1 in mice. Mol Biol Evol 21: 2202-2209.
28. Ross J (1996) Control of messenger RNA stability in higher eukaryotes. Trends Genet 12: 171-175.
29. Pink RC, Carter DRF (2013) Pseudogenes as regulators of biological function. Essays Biochem 54: 103-112.
30. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biology 7: S4.
31. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, et al. (2006) PseudoPipe: An automated pseudogene identification pipeline. Bioinformatics 22: 1437-1439.
32. Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8: S21.
33. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res 31: 3580-3585.
34. Mitrophanov AY, Borodovsky M (2005) Statistical significance in biological sequence analysis. Brief Bioinform 7: 2-24.

35. Durbin R, Eddy SR, Krogh A, Mitchinson G (1998) Biological Sequence Analysis. Cambridge University Press.

36. Gropl C (2012) Markov Chains and Hidden Markov Models.

37. Isaev A (2004) Introduction to Mathematical Methods. Bioinformatics in Springer.

38. Haggstrom O (2002) Finite Markov Chains in Algorithmic Applications. Cambridge University Press.

39. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, et al. (1993) Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. Science 262: 208-214.

40. Rouchka EC (2008) A brief Overview of Gibbs Sampling. University of Louisville Bioinformatics Laboratory Technical Report.

41. Bier A, Oviedo-Landaverde I, Zhao J, Mamane Y, Kandouz M, et al. (2009) Connexin43 pseudogene in breast cancer cells offers a novel therapeutic target. Mol Cancer Ther 8: 786-793.

42. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 8: 534-538.

43. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, et al. (2010) Orphan CpG islands identify numerous conserve promoters in the mammalian genome. PLoS Genet 6: 786-793.

44. Zheng D, Gerstein MB (2006) A computational approach for identifying pseudogenes in the ENCODE regions. Genome Biol 7: S13.

45. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 197 : 261-282.

46. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci USA 99: 3740-3745.

47. Barazandeh A, Mohammadabadi MR, Ghaderi-Zefrehei M, Rafeied F, Imumorin IG, et al. (2019) Whole genome comparative analysis of CpG islands in camelid and other mammalian genomes. Mammalian Biology 98: 73-79.

48. Barazandeh A, Mohammadabadi MR, Ghaderi M, Nezamabadipour H (2016) Genome-wide analysis of CpG islands in some livestock genomes and their relationship with genomic features. Czech Journal of Animal Science 61: 487-495.

49. Barazandeh A, Mohammadabadi MR, Ghaderi-Zefrehei M, Nezamabadipour H (2016) Predicting CpG Islands and Their Relationship with Genomic Feature in Cattle by Hidden Markov Model Algorithm. Iranian Journal of Applied Animal Science 6: 571-579.