Non-coding DNA: A Methodology for Detection and Analysis of Pseudogenes

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

Keywords: Pseudogenes, CpG Island, Alignment, Viterbi Algorithm, Gibbs Sampling.

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 detected 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 strand. These loci can be classed as pseudogene fragments.

1 INTRODUCTION

Completing the human genome reference sequence was a milestone in modern biology. It was quickly recognized that nearly 99% of the $\sim$ 3.3 billion nucleotides that constitute the human genome does not code for proteins (Lander et al., 2001). More recently, studies have discovered many loci that contribute to human diseases and susceptibility to disorders lying outside the protein coding regions (Maurano et al., 2012; Schaub et al., 2012; Martinez et al., 2016; Amiel et al., 2010; Braconi et al., 2011; Bao et al., 2016; Zhang and Zhangm 2015). 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 (Maurano et al., 2012; Schaub et al., 2012; Martinez et al., 2016; Amiel et al., 2010; Braconi et al., 2011; Bao et al., 2016; Zhang and Zhangm 2015).

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 ($\sim$ 20.000) (Koonin, 2005; Zheng et al., 2007). Some pseudogenes are clearly non-functional gene relics (Niimura and Nei, 2007). 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 (lncRNA) transcripts.

Characterizing the pseudogenes and understanding their regulatory role is essential to discover the genetic background of many diseases and to elaborate new pharmacological treatments. Moreover, the correct identification of pseudogenes is important also for gene annotation (Zheng et al., 2007; Zheng and Gerstein, 2006). Despite protein sequence similarity to parent genes is the main feature used to detect pseudogenes, because it is deemed the most sensitive indicator (Harrow et al., 2006; Zhang et al., 2006), we developed an algorithm capable to detect pseudogenes (in particular, processed pseudogenes). Our algorithm 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

Trucco, G. and Cerioli, V.
Non-coding DNA: A Methodology for Detection and Analysis of Pseudogenes.
DOI: 10.5220/0010190400930100
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 3: BIOINFORMATICS*, pages 93-100
ISBN: 978-989-758-490-9

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. 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 (Das and Dai, 2007; Thompson et al., 2003).

2 METHODS

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

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

2.1 Identification of Pseudogene Sequences

The first step is design of a strategy for identifying the pseudogenes of a gene originated 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 $\pm CDS_P \cdot 0.05$ and a nucleotide occurrence in the interval $\pm CDS_{C_n} \cdot 0.2$, for each nucleotide, the extremities 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 (Mitrophanov and Borodovsky, 2005). The parameters of the alignment algorithm are: match = 1, mismatch = 0 and gap = -1.

2.2 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 (Zhang et al., 2002). Processed pseudogenes are typically characterized by the lack of introns and the presence of residue of the polyA tail (Zhang et al., 2002). 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' extremity. 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.

2.3 CpG Islands

CpG islands are regions of DNA in which a cytosine is followed by a guanine in the sequence of nucleotides along the $5' \rightarrow 3'$ direction with a high frequency. The notation CpG is used to distinguish the single strand sequence from the CG pairing on the double strand. 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% (Gardiner-Garden and Frommer, 1987). CpG islands play an important role in gene expression regulation and the ability 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 lengths 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 Hidden Markov Model (HMM). A general HMM (Durbin et al., 1998) is a triplet

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

where:

- Q is an alphabet of symbols;
- S is a finite set of states capable of emitting symbols from the alphabet Q;
- Θ is a set of probabilities, comprised of:
  - state transition probabilities, denoted as p_{ij} for each $i, j \in S$;
  - emission probabilities denoted as $q_k(b)$ for each $k \in S$ and $b \in Q$.

The HMM for CpG islands has (Gröpl, 2012):

- 9 states: begin/end, A+, C+, G+, T+, A-, C-, G- and T-
- 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 (Gröpl, 2012). In this model, each state emits only the corresponding symbol/nucleotide (with probability 1).

The state transition probabilities matrix is reported in Table 1. Model "+" describes the transition probabilities inside the CpG, model "-" describes the transition probabilities outside the CpG island (Gröpl, 2012).

2.4 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 (Haggström, 2002). In bioinformatics, Gibbs sampling is used to detect motif signals in multiple DNA or protein sequences assuming no prior information about the motifs (Das and Dai, 2007; Thompson et al., 2003; Lawrence et al., 1993). Thus, given a set of sequences $S = S^{(1)}, \ldots, S^{(n)}$ 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 (Lawrence et al., 1993; Rouchka, 2008). 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:

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

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 $a^{(1)}, \ldots, s^{(n)}$, the starting indices of the subsequences in $S^{(1)}, \ldots, S^{(n)}$, respectively.
2. Randomly chooses $1 \leq z \leq n$ and computes c_{ij}, qij 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

Table 1: Transition matrix.

	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

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.

3 RESULTS AND DISCUSSION

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 (Zhang et al., 2002). 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 didn't 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 classed as 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 (Zhang et al., 2002).

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

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.

- 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.
- 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.
- The 1000 bp length upstream region of each pseudogene was decoded by the Viterbi algorithm based on a HMM suited for CpG islands detection.
- 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: RPL14, RPL19, RPL22, RPL36 and RPL37 that are ribosomal protein genes (RP family) (Zheng et al., 2007); PTEN (phosphatase and tensin homolog) codes for a tumor suppressor (Chiefari et al., 2010); KRAS (GTPase

Table 2: 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.

	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 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**
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%

KRAS) is a proto-oncogene (Poliseno et al., 2010); RAP1A and RAP1B are members of the oncogene RAS family; 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 (Bier et al., 2009); and finally HDAC1 (histone deacetylase 1) (Tam et al., 2008).

3.1 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.

Table 3 shows the detection results for RPL14. The first column reports the position in the sequence (the 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.

Similarly, we calculated the results for RPL19, RPL22, and RPL37, which are not reported here due to lack of space.

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 4 reports the computation time of each experiment.

3.2 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 5 shows the number of tails and the number of polyadenylation signals found for each group of pseudogenes.

3.3 CpG Islands

The upstream 1000 bp length regions of the detected pseudogenes were analyzed in order to check the presence of CpG islands. Table 6 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

Table 4: The table shows the length of each CDS and the computation time needed to scan the entire genome in search of its pseudogenes.

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 5: 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.

	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**

varies a lot and, therefore, a sliding window cannot be used to detect CpG islands.

3.4 Motif Discovery and Flanking Regions

It was observed that half of mammalian CpG islands ($\sim$ 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 (Illingworth et al., 2010). 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 did not 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' extremities. 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 didn't test the ribosomal pseudogenes for this feature.

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 7, 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.

4 CONCLUSIONS

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 (Pink and Carter, 2013). Characterizing the pseudogenes and understanding their regulatory role will help in discovering the genetic origin of many diseases but also in finding 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 (Zheng et al., 2007; Zheng and Gerstein, 2006). 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 (Zheng et al., 2007). The algorithm based on raw nucleotide identity, even if it did 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 (Zhang et al., 2002). 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

Table 6: 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.

	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 7: 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.

	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%	

detected (Zheng et al., 2007).

PolyA tails. Although it was observed that a polyA tail is present beyond a processed pseudogene in about half of the cases (Zhang et al., 2002), 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 (Gardiner-Garden and Frommer, 1987). However, any definition of CpG island, after all, is arbitrary (Takai and Jones, 2002). 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.

REFERENCES

E. S. Lander et al., International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome, in *Nature* 409 (2001), 860-921, doi: 10.1038/35057062

M. T. Maurano et al. Systematic localization of common disease-associated variation in regulatory DNA. in *Science* 337 (2012), 1190-1195. doi: 10.1126/science.1222794

M. A. Schaub et al. Linking disease associations with regulatory information in the human genome. in *Genome Research* 22(9) (2012), 1748-2759. doi: 10.1101/gr.136127.111

A. F. Martinez et al. An ultraconserved brain-specific enhancer within DGRL3 (LPHN3) underpins attention-deficit/hyperactivity disorder susceptibility. in *Biological Psychiatry* 80 (2016), 943-954. doi: 10.1016/j.biopsych.2016.06.026

J. Amiel, S. Benko, C.T. Gordon and S. Lyonnet. Disruption of long-distance higly conserved noncoding elements in neurocristopathies. in *Annals of the New York Academy Sciences* 1214 (2010), 34-46

C. Braconi et al. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. in *Proceedings of the National Academy of Sciences* 108 (2011), 786-791. doi: 10.1073/pnas.1010198108.

B. Bao et al. Genetic variants in ultraconserved regions associate with prostate cancer recurrence and survival. in *Scientific Reports* 6 (2016), 22124 doi: 10.1038/srep22124

Feng Zhang and James R. Zhang. Non-coding genetic variants in human disease. in *Human Molecular Genetics* 24 (2015), R102-R110. doi: 10.1093/hmg/ddv259

Eugene V. Koonin. Orthologs, Paralogs and Evolutionary Genomics. in *Annual Review of Genetics* 39(1) (2005), 309-338. doi: 10.1146/annurev.genet.39.073003.114725

D. Zheng et al. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. in *Genome Research* 17 (2007), 839-851. doi: 10.1101/gr.5586307

Yoshihito Niimura, Masatoshi Nei. Extensive gains and losses of olfactory receptor genes in mammalian evolution. in *PLoS ONE* 2(8) (2007), 860-921. doi: 10.1371/journal.pone.0000708

Zhaolei Zhang, Paul Harrison, Mark Gerstein. Identification and analysis of over 200 ribosomal protein pseudogenes in the human genome. in *Genome Research* 12(10) (2002), 1466-1482. doi:10.1101/gr.331902

L. Poliseno et al. A coding-independent function of gene and pseudogene mRNAs regulates tumor biology. in *Nature* 465 (2010), 1033-1038. doi: 10.1038/nature09144

E. Chiefari et al. Pseudogene-mediated pstrascriptional silencing of HMGA1 can result in insulin resistance and Type 2 diabetes. in *Nature Communications* 12(10) (2010), 1-7. doi: 10.1038/ncomms1040

Ryan C. Pink, David R.F Carter. Pseudogenes as regulators of biological function. in *Essays in Biochemestry* 54 (2013), 103-112. doi: 10.1042/bse0540103

Jennifer Harrow, France Denoeud, Adam Frankish et al. GENCODE: producing a reference annotation for ENCODE. in *Genome Biology* 7 (2006), S4. doi: 10.1186/gb-2006-7-s1-s4

Z. Zhang et al. PseudoPipe: an automated pseudogene identification pipeline. in *Bioinformatics* 22 (2006), 1437-1439. doi: 10.1093/bioinformatics/btl116

Modan K. Das and Ho-Kwok Dai. A survey of DNA motif finding algorithms. in *BMC Bioinformatics* 8(7) (2007), S21. doi: 10.1186/1471-2105-8-S7-S21.

William Thompson, Eric C. Rouchka and Charles E. Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. in *Nucleic Acid Research* 31(13) (2003), 3580-3585. doi: 10.1093/nar/gkg608

Alexander Yu Mitrophanov and Mark Borodovsky. Statistical significance in biological sequence analysis. in *Briefings in Bioinformatics* 7(1) (2005), 2-24. doi: 10.1093/bib/bbk001

Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchinson. Biological Sequence Analysis. in *Cambridge University Press* (1998).

Clemens Gröpl. Markov Chains and Hidden Markov Models in *www.mi.fu-berlin.de* (2012)

Olle Haggström. Finite Markov Chains in Algorithmic Applications in *Cambridge University Press* (2002)

Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, John C. Wootton. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. in *Science* 262 (1993), 208-214. doi: 10.1126/science.8211139

Eric C. Rouchka. A brief Overview of Gibbs Sampling. in *University of Louisville Bioinformatics Laboratory Technical Report* 02 (2008).

A. Bier et al. Connexin43 pseudogene in breast cancer cells offers a novel therapeutic target. in *Molecular Cancer Therapeutics* 8(4) (2009), 786-793. doi: 10.1158/1535-7163.MCT-08-0930

Oliver H. Tam, Alaxei A. Aravin, Paula Stein et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. in *Nature* 8(4) (2008), 534-538. doi: 10.1038/nature06904

R. S. Illingworth et al. Orphan CpG islands identify numerous conserve promoters in the mammalian genome. in *Plos Genetics* 6 (2010), 786-793. doi: 10.1371/journal.pgen.1001134

Deyou Zheng and Mark B. Gerstein. A computational approach for identifying pseudogenes in the ENCODE regions. in *Genome Biology* 7 (2006), S13. doi: 10.1186/gb-2006-7-s1-s13

Margaret Gardiner-Garden and Marianne Frommer. CpG islands in vertebrate genomes. in *Journal of Molecular Biology* 197 (1987), 261-282. doi: 10.1016/022-2836(87)90689-9

Daiya Takai, Peter A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. in *Proceedings of the National Academy of Sciences* 99(6) (2002), 3740-3745. doi: 10.1073/pnas/05240099