PseudoDomain: identification of processed pseudogenes based on protein domain classification

Yuan Zhang
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48824, U.S.A
zhangy72@msu.edu

Yanni Sun
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48824, U.S.A
yannisun@msu.edu

ABSTRACT

Pseudogenes are dysfunctional DNA sequences that share sequence similarities with functional genes. Accurate identification of pseudogenes is important to understand biological and evolutionary histories of genomes and genes. Most existing pseudogene identification tools rely on homology search between genomic sequences and annotated proteins of the same genome. However, when accurate annotations of the genome of interest are not available, these tools will not be able to provide reliable pseudogene identification. In this work, we introduce a new pseudogene identification tool named PseudoDomain, which is designed to accurately identify processed pseudogenes in genomes with or without gene annotations. PseudoDomain uses profile Hidden Markov Model-based homology search between genomic sequences and protein domain families, which are conserved in a large number of proteins. Experimental results show that our method is able to effectively identify processed pseudogenes with high sensitivity and low false positive rate. In addition, it can accurately predict the number and positions of frameshifts within putative pseudogenes. The source codes of PseudoDomain are available at `http://sourceforge.net/projects/pseudodomain/`.

Categories and Subject Descriptors

J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics

General Terms

Application

Keywords

Pseudogene identification, protein domain classification, profile HMM, frameshift detection

1. INTRODUCTION

ACM-BCB'12, October 7-10, 2012, Orlando, FL, USA

Pseudogenes are complete or partial copies of genes and can no longer encode functional gene products [1, 2]. There are a large number of pseudogenes in mammals such as human and mouse. Pseudogenes are genomic fossils that are important resources for the study of evolutionary histories of particular genes or gene families. For example, human type I hair keratin pseudogene phihHaA is found to be differentially expressed in chimpanzee and gorilla, showing the recent inactivation of the human gene after the Pan-Homo divergence [3]. Pseudogenes can also be used to determine different forms and rates of neutral sequence evolution among different regions in the genome and even among different organisms [4]. Some pseudogenes are reported to be transcribed and even be functional [5]. It is found that an expressed pseudogene regulates the stability of messenger-RNA of its parent coding gene [6].

In the past years, several groups have made intensive efforts to identity pseudogenes in mammalian genomes. In an initial sequencing and comparative analysis of the mouse genome, it is reported that there are about 14,000 putative pseudogenes [7]. Torrents et al. [4] conducted a genome-wide survey of human pseudogenes and identified about 20,000 pseudogenes from all intergenic regions in the human genome. They further estimated that these pseudogenes only accounted for a small fraction of the total number of the pseudogenes in the human genome. In another study Zhang et al. [8] reported 8,000–12,000 pseudogenes in the human genome and 5,000 in the mouse genome .

Pseudogenes can generally be classified into non-processed pseudogenes and processed psuedogenes. Non-processed pseudogenes are generated by gene duplication events and are also called duplicated pseudogenes. They usually keep the original exon-intron structures of their parent genes. Processed pseudogenes are generated by random insertions of mature mRNAs back into the genome. Processed pseudogenes usually lack exon-intron structures. Pseudogenes are generally unconstrained by selection pressure and accumulate mutations, leading to frame disruptions such as stop codons, frameshifts or interspersed repeats [9].

There are more processed pseudogenes than non-processed pseudogenes in the human genome [4, 8]. It is estimated that there are 9,000–11,000 processed pseudogenes in the human genome [10]. Zhang et al. [8] identified about 8000 processed pseudogenes in the human genome. The identi-

fication of processed pseudogenes provides important information of the rate and age of retrotransposition events [4]. Processed pseudogenes can also serve as fossilized footprint of the expression of their parent genes [11]. Thus, in this work, we mainly focus on identifying processed pseudogenes in the human genome.

Most current pseudogene identification tools use BLAST-based [12] homology searches between intergenic regions and annotated proteins of the same genome, which are retrieved from known protein databases such as the RefSeq protein database [13] and the SP-TREMBL database [14]. However, these tools have some limitations. First, they rely on the annotations of the genome. When reliable annotations are not available, the errors in genome annotations will be inherited by pseudogene annotation. Although the development of next-generation sequencing (NGS) technologies has greatly improved annotations of genomes [15], systematic errors of NGS technologies [16] as well as intrinsic errors in genome annotations [17] still impose great challenges for accurate genome annotations. For newly sequenced genomes that have not been carefully annotated, this method is not able to provide accurate pseudogene annotation. Second, some pseudogenes diverge from their parent genes due to accumulation of mutations. These pseudogenes may be missed by BLAST-based homology searches.

In order to address these challenges, we introduce a processed pseudogene identification tool, PseudoDomain, based on protein domain classification. A protein domain is a structural and functional unit that is independent of the rest of the protein molecule. A protein may have a single or multiple domains as shown in Figure 1.

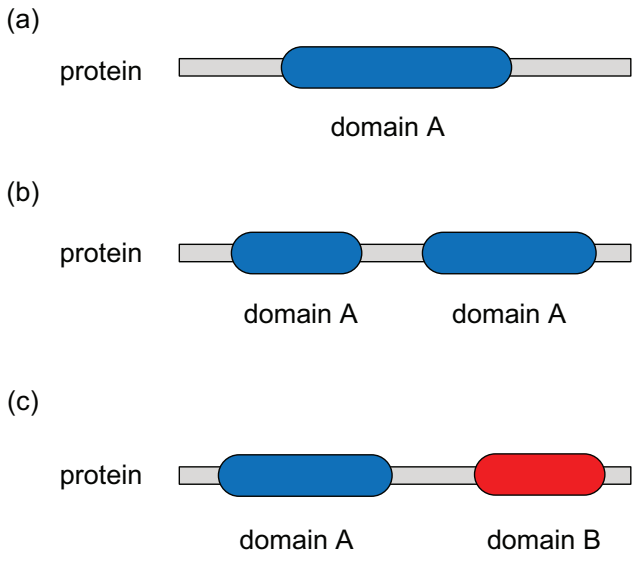

Figure 1: Different domain organizations in three different proteins. (a) A protein with a single domain. (b) A protein with multiple copies of the same domain. (c) A protein with multiple domains.

The identification of protein domains helps us understand the evolution, structure and function of protein families. Pfam [18] protein domain database has a large collection of annotated protein domain families, which are built on sets of homologous protein regions that share significant sequence similarities. Its latest version, Pfam 26.0 consisits of about 13,000 Pfam-A families, which are manually curated families with high qualities. Pfam uses UniProt Knowledgebase (UniProtKB) [19] as its reference sequence database. Its coverage of UniProtKB is currently as high as about 80% [18]. Pfam also has high coverage of the proteins in eukaryotic genomes such as the human and mouse genomes. Moreover, Pfam keeps a fast-paced growth with new families being added and copes well with the increase of the number of protein sequences in protein databases. The high coverage of domain families in existing proteins builds the foundation for pseudogene identification without genome annotations. When gene duplication or retroposition events happen, protein domains of parent genes are copied to pseudogenes. Therefore, even when protein annotations of a mammalian genome are unavailable, putative pseudogenes from this genome are still likely to be classified into protein domain families annotated in Pfam. PseudoDomain takes the advantage of Pfam's high coverage and can accurately identify protein domains of processed pseudogenes without relying on genome annotations. The identification of protein domains indicates the existence of pseudogenes or protein coding genes. Based on the absence of introns and presence of frame disruptions whithin processed pseudogenes, PseudoDomain can successfully distinguish processed pseudogenes from protein coding genes.

HMMER [20] is a profile Hidden Markov Model (HMM)-based protein domain classification tool. In conjunction with the Pfam database, where each protein domain family is represented by a profile HMM, HMMER can search sequence databases for protein sequences that are homologous to annotated protein domain families. PseudoDomain uses HMMER to search query genomic sequences for regions that share similarities with annotated protein domain families in Pfam-A and classifies these regions to their corresponding families. Its latest version, HMMER 3.0 has achieved comparable speed to BLAST, making it suitable to analyze large-scale data sets [21]. This enables PseudoDomain to obtain high efficiency in identifying processed pseudogenes in large mammalian genomes. Because profile HMM-based homology search has high sensitivity in classifying remote homologs [22], PseudoDomain is more sensitive in identifying putative pseudogenes with intensive mutations. Moreover, each genomic sequence only needs to be searched against all protein domain families, the number of which is generally smaller than annotated proteins of most mammalian genomes. In addition, PseudoDomain uses HMM-FRAME [23] to automatically detect the number and positions of frameshifts within the pseudogene sequences. This provides important evidence for pseudogene identification.

2. RELATED WORK

A number of methods have been proposed to identify pseudogenes in mammalian genomes. PseudoPipe [24] and PseudoGeneQuest [25] both search all intergenic regions for hits that share sequence similarities with annotated proteins in the genome. BLAST hits are then processed to form the final set of predicted pseudogenes. Torrents et al. [4] screened all intergenic regions in the human genome to identify pseu-

dogenes with a combination of homology search and a functionality test using the ratio of silent to replacement nucleotide substitutions. Zheng et al. [26] proposed a computational pipeline to explicitly use exon-intron structures to classify pseudogenes. It can also be used to distinguish between duplicated and processed pseudogenes. Although some of these methods can identify pseudogenes and classify non-processed pseudogenes and processed pseudogenes, all of them rely on homology searches between query sequences and annotated proteins. This work provides a complementary pseudogene identification method for genomes lacking quality annotations.

3. METHOD

3.1 Pipeline of PseudoDomain

In order to identify processed pseudogenes, we make use of their following features: 1) most processed pseudogenes can be classified into protein domain families, 2) processed pseudogenes typically have frame disruptions such as frameshifts and stop codons, 3) processed pseudogenes have their introns spliced out. The first feature enables us to extract a majority of genes and pseudogenes from genomic sequences. The other two features enable us to distinguish processed pseudogenes from protein coding genes.

PseudoDomain incorporates these features and can be divided into five main steps accordingly: profile HMM-based protein domain classification, elimination of redundant hits, clustering of neighboring hits, domain coverage filtration, and analysis of frame disruptions. In the first step, we use HMMER to obtain all genomic regions that are found to have significant similarities with some protein domain families as well as the alignments between these regions and corresponding protein domain families. These regions are called *raw hits*. In the second step, we eliminate hits that have significant overlaps with other hits with better E-values. This step helps us remove random matches and determine the original protein domain families of raw hits. In the third step, we use a procedure called ClusteringNeighboringHits to cluster neighboring hits belonging to the same processed pseudogene structure. Segmented hits of the same processed pseudogenes will be merged to form non-redundant hits. In the forth step, we use domain coverage filtration to eliminate hits generated by random matches. In the last step, we analyze frame disruptions in our predicted processed pseudogenes including frameshifts and stop codons. Frame disruption information is helpful to verify our prediction and analyze factors that lead to the dysfunctionality of these pseudogenes. Figure 2 shows a schematic representation of the pipeline of PseudoDomain. The inputs of the pipeline are genomic sequences and protein domain families. The outputs of the pipeline are putative processed pseudogene regions in the input genomic sequences and frame disruption annotations.

3.2 Protein domain classification for genomic sequences

In this step, we use HMMER to classify genomic sequences into annotated protein domain families. HMMER is used to search protein domain databases for homologs of protein sequences. We first translate genomic sequences into peptide sequences using 6-frame translations. We then search all

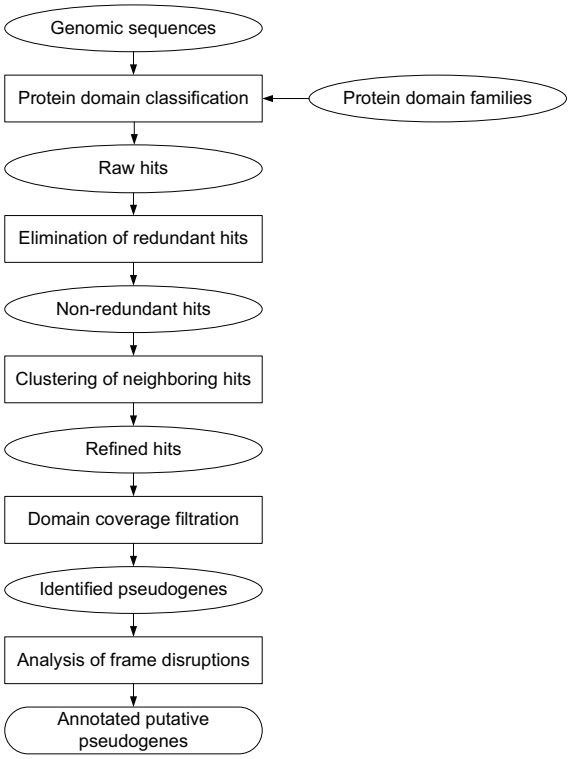

Figure 2: The pipeline of PseudoDomain

13,672 protein domain families in Pfam-A against the translated sequences using gathering thresholds (GAs). GAs are chosen with the goal of maximizing coverage while excluding any false positive matches [18]. HMMER indicates beginning and ending positions of raw hits in genomic sequences as well as protein domain families. This information will be essential for us to eliminate redundancies and cluster neighboring hits.

3.3 Elimination of redundant hits

In this step, we try to eliminate redundancies of raw hits. Redundant hits are defined as hits that have significant overlaps with other hits that have better E-values. These redundancies arise generally in two cases. In one case, some genomic sequences are classified into multiple protein domain families. In the other case, overlapping parts of the same genomic region are classified into different parts of a single protein domain family. In both cases, we eliminate hits that have worse E-values. Here, a significant overlap is determined by the length of the overlapped region of two neighboring hits. If this value is larger than half of the length of the shorter hit, these two hits are defined to have a significant overlap.

3.4 Clustering of neighboring hits

Since processed pseudogenes do not have introns, each hit represents a domain region from a processed pseudogene that can be classified into a protein domain family. However, pseudogenes often have highly diverged regions or frame disruptions within their domain regions. Therefore, hits from the same pseudogene region tend to be segmented due to low

sequence similarities in these regions. For processed pseudogenes that have multiple domains, they tend to be close to each other due to the absence of introns. Figure 3 shows different distributions of protein domains in parent genes and their processed pseudogenes. This comparison helps us distinguish genes from processed pseudogenes. For convenience of explanation, the influence of frameshifts is not indicated. However, this does not affect the patterns of protein domain distributions.

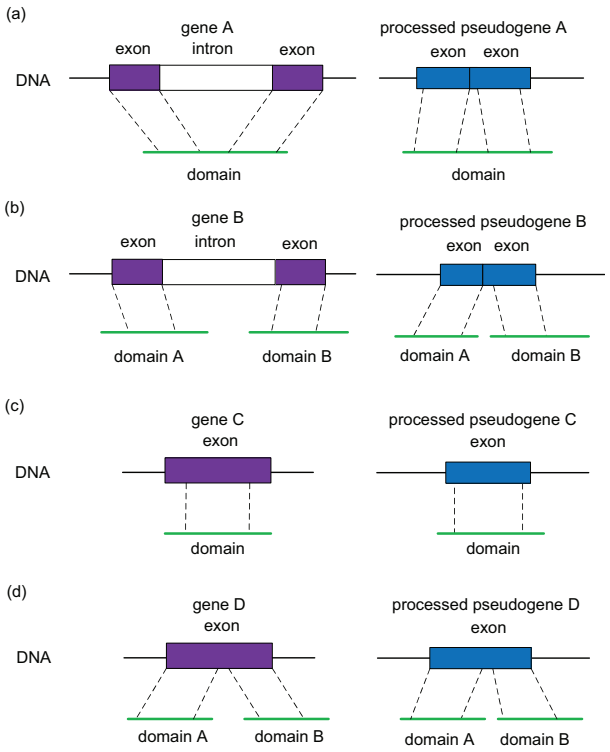

Figure 3: A comparison of domain distributions in parent genes and their processed pseudogenes. (a) In both parent gene A and processed pseudogene A, two exons contain two regions of a single protein domain. (b) In both parent gene A and processed pseudogene A, two exons contain two different protein domains. (c) In both parent gene A and processed pseudogene A, a single exon contains a single protein domain. (d) In both parent gene A and processed pseudogene A, a single exon contains two different protein domains

In Figure 3, although parent gene A and processed pseudogene A both have segmented regions that can be classified by HMMER, the distance of two hits from gene A tend to be much larger than that of two hits from pseudogene A. For the first two cases in Figure 3, we can apply a distance threshold to distinguish genes from processed pseudogenes. For the last two cases, protein domain distributions of both parent genes and processed pseudogenes are very similar. When we do not know gene annotations, it is difficult to distinguish between hits from parent genes and processed pseudogenes. However, introns widely exist in protein coding genes of mammalian genomes. On average, the number of introns per gene is 7.8 for human [27]. In most cases, protein domains exist in multiple exons of genes. Therefore, PseudoDomain can maintain high accuracy in identifying processed pseudogenes in genomes that have not been annotated.

In the human genome, the average intron size is about 5419 bp with less than 0.01% of the introns smaller than 20 bp and more than 90% introns larger than 60 bp in length [28]. In the mouse genome, there are also about 90% introns with larger than 60 bp in length. Thus, we use 60 bp as the distance threshold of non-redundant hits in order to distinguish processed pseudogenes from protein coding genes.

Based on these observations, we introduce a procedure called ClusterNeighboringHits to cluster neighboring hits that belong to the same pseudogene structure. We aim at producing a refined hit set in which regions from the same processed pseudogene structures are clustered and merged.

Let $U = \{U_1, U_2, ..., U_N\}$ denote the input set of non-redundant hits. Let V denote the output set of refined hits. Let $N = |U|$ denote the number of hits in U. Let $U_i.begin$ and $U_i.end$ denote the beginning and ending positions of hits in the query genomic sequence. For the convenience of merging neighboring hits, hits in U are sorted according to their beginning positions in the genomic sequence. Let $D_{i,j}$ denote the distance between U_i and U_j in the genomic sequence, where $i \leq j$. The distance between two hits is defined as the distance between the nearest positions of the two hits. If two hits have overlaps their distance is defined as zero. Equation 1 gives a formal definition of distance between U_i and U_j.

$$D_{i,j} = \begin{cases} U_j.begin - U_i.end - 1 & \text{if } U_j.begin > U_i.end \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Let D^* denote the distance threshold between two neighboring hits in the genomic sequence. Neighboring hits whose distance in the genomic sequence is no larger than D^* will be merged. This distance threshold is critical for us to avoid classifying protein coding genes as processed pseudogenes. The default value of D^* is 60 bp, which is chosen based on our previous discussion. Procedure 1 shows the pseudocode of ClusterNeighboringHits procedure. Figure 4 shows an example of how PseudoDomain merges four hits from the same protein domain family.

3.5 Domain coverage filtration

In this step, we eliminate non-redundant hits that have low domain coverage. Domain coverage is defined as the fraction of positions covered by a non-redundant hit in a protein domain. In the human genome, more than 80% of protein coding genes have a protein domain coverage of at least 50%. Therefore, it is more likely that hits of very low domain coverage are random matches. Segmented hits from a processed pseudogene structure are usually clustered into a single hit. This hit preserves the original domain region of the processed pseudogene and has relatively high coverage of the protein domain. Therefore, we use domain coverage filtration to remove all refined hits that do not belong to pro-

Procedure 1 ClusterNeighboringHits

Input: U: a set of N non-redundant hits.
Output: V: a set of refined hits.

```
    //k: index of the first one of two neighboring hits we
    are trying to merge.
 1: k ⇐ 1
 2: for i ⇐ 2 to N do
 3:    if D_{k,i} ≤ D* then
 4:       U_i.begin = U_k.begin
 5:    else
 6:       add U_k into V;
 7:    end if
 8:    k ⇐ i
 9: end for
    //add the last hit.
10: add U_k into V;
```

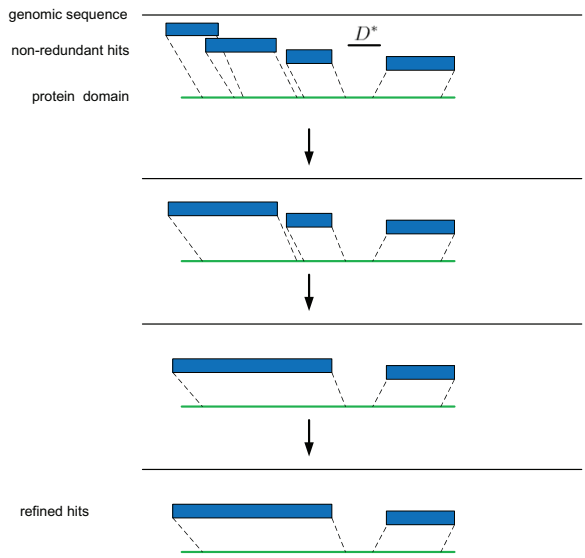

Figure 4: An example of clustering four hits from the same protein domain family.

cessed pseudogenes. Here we apply a user-defined threshold for domain coverage. Refined hits that have domian coverage lower than this threshold will be eliminated. By default, the domain coverage threshold PseudoDomain uses is 50%.

3.6 Analysis of frame disruptions

Here we conduct an analysis of frameshifts and stop codons within pseudogenes we identify. It provides users with better understanding of the factors that lead to the dysfunctionality of processed pseudogenes. HMM-FRAME [23] is originally designed to predict and correct frameshift errors introduced by pyrosequencing technology. Sequence reads generated by pyrosequencing technology usually have much higher frameshift error rates in homopolymer regions than in non-homopolymer regions [29]. HMM-FRAME accepts a sequencing error model as input to accommodate different error rates for sequence reads generated by different platforms. HMM-FRAME is proven to have high sensitivity and accuracy in locating frameshifts using profile-HMM based homology search while maintaining a low false positive rate (FP rate). Frameshifts tend to occur more frequently in pseudogenes than in pyrosequencing reads. We therefore adjust the error model of HMM-FRAME to accommodate this need. The default error rate we use in PsuedoDomain is 0.044. PseudoDomain also detects stop codons within our identified processed pseudogenes. Annotations of frameshifts and stop codons are output by PseudoDomain so that users will be able to use them for future analysis and research.

3.7 Running time analysis

The most time-consuming part of PseudoDomain is protein domain classification by HMMER. HMMER adopts profile HMM Forward/Backward algorithms, whose original time complexity is $O(LM)$ to search genomic sequences against one protein domain family, where L is the total size of the genomic sequences and M is the number of match states in the protein domain family. The latest version of HMMER introduces several acceleration heuristics that make HMMER as fast as BLAST for protein searches [21]. Currently, the shortest protein domain family in Pfam-A has 7 match states and the longest one has 2207. There are 13,672 protein domain families in Pfam-A. The second, third, and forth steps of PseudoDomain have linear time complexity in terms of the total number of hits. In the last step, HMM-FRAME has the time complexity of $O(LM)$ for each processed pseudogene. The size of identified processed pseudogenes is much smaller than that of input genomic sequences. Moreover, each processed pseudogene only needs to be searched against the protein domain families HMMER classifies. Therefore, the running time of PseudoDomain is mainly decided by HMMER.

4. EXPERIMENTAL RESULTS

In order to evaluate the ability of PseudoDomain to accurately identify pseudogenes, we applied PseudoDomain to two data sets. The first one contains processed pseudogenes of the human genome. These pseudogenes are annotated by the PseudoFam [30] database, which contains pseudogenes identified from 10 eukaryotic genomes. These pseudogenes are assigned to different protein domain families in Pfam. As the annotations of these pseudogenes are available, we can quantify the sensitivity of PseudoDomain in this data set. The second data set consists of genomic sequences of all chromosomes of the human genome. We demonstrate the utility of PseudoDomain in identifying processed pseudogenes without using any existing annotations.

4.1 Identification of processed pseudogenes in an annotated data set

In this experiment, we applied PseudoDomain to identify all processed pseudogenes of the human genome that are annotated by PseudoFam. There are totally 7069 pseudogenes in the human genome with 5610 processed pseudogenes and 1459 non-processed pseudogenes. Among the 5610 processed pseudogenes, 4369 have a single protein domain and 1241 have multiple protein domains. We downloaded all the processed pseudogene sequences from the UCSC browser [31] according to their locations in the genome provided by PseudoFam. We also downloaded the Pfam-A data set from the Pfam website [32].

In our first step, we searched all the processed pseudogenes

against all protein domain families in Pfam-A. HMMER produced 7566 raw hits from 4794 processed pseudogene sequences and 912 protein domain families. 85.45% of processed pseudogenes in PseudoFam have significant homologs to protein domain families in Pfam. This further shows that protein domain classification is useful in identifying processed pseudogenes. After we eliminated redundancies in the raw hits and applied ClusteringNeighboringHits procedure, we obtained a set of 4794 refined hits. This shows that a significant number of hits have been clustered. After domain coverage filtration using 50% as the threshold, we finally identified 4341 processed pseudogenes, which accounted for 77.38% of all processed pseudogenes.

There are two main reasons that 22.62% of processed pseudogenes in PseudoFam cannot be identified by PseudoDomain. First, although some pseudogenes share sequence similarities with functional genes, protein domains in their parent genes are not copied to these pseudogenes. In sequence level, BLAST hits between these pseudogenes and their parent genes do not cover domain regions as shown in Figure 5. Second, accumulation of mutations leads to high divergence of protein domain regions in pseudogenes. These regions may not be able to be classified into annotated protein domain families. If we decrease the threshold of domain coverage the sensitivity of PseudoDomain will be increased. However, this will result in more false positive predictions. Therefore, the threshold of 50% domain coverage is recommended so that PseudoDomain will generate a reliable output set of processed pseudogenes. However, users still can specify this parameter to accommodate variations of different data sets.

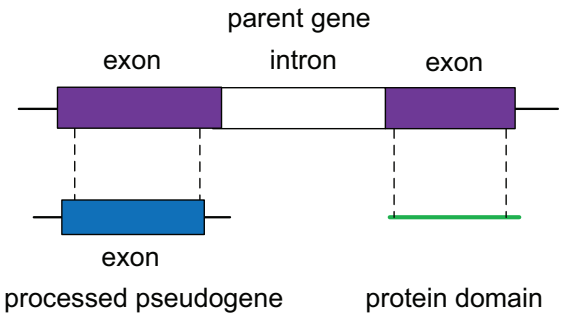

Figure 5: The processed pseudogene and a protein domain share sequence similarities with different parts of the parent gene.

PseudoDomain further detected that 69.32% of identified processed pseudogenes have at least one stop codon. This reveals that stop codons are important factors that lead to the loss of protein coding functionality of processed pseudogenes. Moreover, HMM-FRAME found that 52.88% of identified processed pseudogenes have frameshifts in their protein domain regions. On average, each of these processed pseudogenes has 1.08 frameshifts.

4.2 Annotation of processed pseudogenes in the human genome

In this experiment, we applied PseudoDomain to annotate processed pseudogenes in the human genome (hg19). We downloaded chromosome sequences and genome annotation data from the UCSC genome browser [31]. These sequences were masked by ReapeatMasker [33]. We first used PseudoDomain to search for processed pseudogenes without the knowledge of genome annotations. The parameter settings are identical to those used in the first experiment. While most present work only evaluates FP rate on protein coding genes, we also consider other genomic features. We compared our identified processed pseudogenes (denoted as M) with annotated genes including protein coding genes, tRNA genes, and sno/miRNA genes (denoted as N). FP rate is defined as $\frac{M \cap N}{N}$. Table 1 shows statistical features of processed pseudogenes in each chromosome output by PseudoDomain.

From Table 1 we can see that the FP rates of PseudoDomain are very low in all chromosomes. These false positive cases are mainly protein coding genes that are classified as processed pseudogenes by PseudoDomain. Most of these false positive cases fall into the last two cases as shown in Figure 3. Only two ncRNA genes are classified as processed pseudogenes in all chromosomes. These results show that PseudoDomain can accurately distinguish other genomic features from processed pseudogenes.

The average number of stop codons and frameshifts in each processed pseudogenes indicates the frequency of frame disruptions. Although HMM-FRAME can accurately locate and correct frameshifts within sequences that are homologous to protein domain families, it cannot handle frameshifts generated by continuous mutations in highly divergent regions of the query sequences. Therefore, the average number of frameshifts in each processed pseudogene should be actually larger than that output by HMM-FRAME.

This experiment was run on a 2.2 GHz dual-core AMD Opteron machine. We ran HMMER on all the 24 chromosomes concurrently. For each chromosome, each of the 6 frames of the translated peptide sequences can also be run concurrently. On average, it took HMMER around 2 hours to search one translated peptide sequence in each frame against all protein domain families. After HMMER generated all the raw hits, the remaining steps totally took less than one hour for all chromosomes. PseudoPipe is reported to take approximately one day on a machine of similar configurations to finish genomic sequences of a comparable size of one chromosome in the human genome. PseudoDomain ran significantly faster than PseudoPipe.

5. CONCLUSION AND FUTURE WORK

In this work, we introduce PseudoDomain, a processed pseudogene identification tool based on protein domain classification. It provides accurate identification of procssed pseudogenes by searching genomic sequences for homologs to annotated protein domain families in Pfam. Unlike existing pseudogene identfication tools, it does not rely on annotations of genomes. Therefore, for newly sequenced genomes PseudoDomain is able to conduct annotations of processed pseudogenes. Experimental results show that it successfully distinguishes between putative processed pseudogenes and protein coding genes and achieves a good trade-off between sensitivity and FP rate. As our tool relies on annotated

Table 1: Features of processed pseudogenes identified by PseudoDomain

Chromosome	Number	Average length	Average stop codon number	Average frameshifts	FP rate
chr1	769	368.83	1.41	0.95	2.72%
chr2	490	403.75	2.37	0.97	2.28%
chr3	330	424.62	2.58	1.16	2.65%
chr4	331	468.62	2.89	0.64	2.76%
chr5	311	406.42	2.01	1.10	5.35%
chr6	386	399.25	1.77	0.96	2.70%
chr7	367	386.47	2.13	0.95	1.02%
chr8	240	423.86	2.30	1.12	2.57%
chr9	265	451.54	2.50	1.08	1.01%
chr10	262	375.27	2.02	0.85	1.15%
chr11	335	444.82	1.76	0.54	2.51%
chr12	302	402.52	2.44	0.86	1.56%
chr13	140	412.89	2.26	1.18	3.42%
chr14	338	382.24	1.38	0.87	2.99%
chr15	201	377.97	1.33	0.74	1.06%
chr16	184	476.3	3.54	0.29	3.12%
chr17	168	385.29	1.66	0.52	2.38%
chr18	91	377.67	2.00	1.00	0.84%
chr19	226	385.43	1.81	0.78	2.36%
chr20	98	374.85	2.13	0.86	1.42%
chr21	50	432.96	2.64	1.14	2.03%
chr22	147	405.94	3.31	0.91	2.26%
chrX	398	471.73	2.90	0.82	4.77%
chrY	47	453.19	3.70	0.41	3.57%

protein domains for pseudogene annotation, it cannot conveniently reveal the parent genes of the pseudogenes. We plan to use the comparison of the domain organization to relate the parent genes to the pseudogenes. Moreover, it currently only identifies processed pseudogenes. Protein domain classification is useful in identifying putative genomic sequences that share sequence similarity with exiting protein domain families. In our future work, we will explore better distinguishing features of non-processed pseudogenes so that PseudoDomain will be able to provide comprehensive and accurate pseudogene identifications. We will also conduct statistical analysis of the different distributions of distance of protein domains in genes and pseudogenes. This will help us improve accuracy of pseudogene identification.

6. ACKNOWLEDGEMENTS

This work was partially supported by the NSF grants DBI-0953738.

7. REFERENCES

[1] E. Vanin. Processed pseudogenes: Characteristics and evolution. *Hum Genet.*, 19:253–272, December 1985.

[2] A.J. Mighell and N.R. Smith and P.A. Robinson and A.F. Markham. Vertebrate pseudogenese. *FEBS Lett.*, 268(2-3):109–14, February 2000.

[3] H. Winter et al. Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum Genet.*, 108(1):37–42, January 2001.

[4] D. Torrents and M. Suyama and E. Suyama and P. Bork. A genome-wide survey of human pseudogenes. *Genome Res*, 13(12):2559–67, December 2003.

[5] E. S. Balakirev and F. J. Ayala. Pseudogenes: Are they "Junk" or functional DNA? *Annu Rev Genet*, 37:123–151, December 2003.

[6] S. Hirotsune et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, 423(6935):91–6, May 2003.

[7] R. H. Waterston et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, December 2002.

[8] Z. Zhang, N. Carriero, and M. Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.*, 20(2):62–7, Feburary 2004.

[9] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, February 1968.

[10] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, Feburary 2001.

[11] I. Podlaha and J. Zhang. Processed pseudogenes: the 'fossilized footprints' of past gene expression. *Trends Genet.*, 25(10):429–434, Feburary 2009.

[12] S. F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, 25(17):3389–3402, September 1997.

[13] NCBI reference sequences. `http://www.ncbi.nlm.nih.gov/RefSeq/`.

[14] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucl. Acids Res.*, 25:31–36, Janurary 1997.

[15] M. Pop and S. L. Salzberg. Bioinformatics challenges of new sequencing technology. *Trends Genet.*, 24(3):142–149, March 2008.

[16] O. Harismendy et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10:R32, March 2009.

[17] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429–31, August 2001.

[18] M. Punta et al. The Pfam protein families database. *Nucl. Acids Res.*, 40(D1):D290–D301, November 2011.

[19] Protein Knowledgebase: UniProtKB. `http://www.uniprot.org/`.

[20] HMMER3: a new generation of sequence homology search software. `http://hmmer.janelia.org/`.

[21] S. R. Eddy. Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, 7:e1002195, October 2011.

[22] K. Karplus and C. Barrett and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–56, October 1998.

[23] Y. Zhang and Y. Sun. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics*, 12(1):198, May.

[24] Z. Zhang et al. Pseudopipe: an automated pseudogene identification pipeline. *Bioinformatics*, 22(12):1437–9, June 2006.

[25] C. Ortutay and C. Vihinen. Pseudogenequest - service for identification of different pseudogene types in the human genome. *BMC Bioinformatics*, 9:299, July 2008.

[26] D. Zheng and M. B. Gerstein. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biology*, 7(suppl 1):S13, August 2006.

[27] M. K. Sakharkar, V. T. Chow, and P. Kangueane. Distributions of exons and introns in the human genome. *In Silico Biol*, 4(4):387–93, 2004.

[28] M. K. Sakharkar et al. An analysis on gene architecture in human and mouse genomes. *In Silico Biol*, 5(4):347–65, May 2005.

[29] C. Wang et al. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.*, 17(8):1195–201, March 2007.

[30] H. Y. K. Lam et al. Pseudofam: the pseudogene families database. *Nucl. Acids Res.*, 37(suppl 1):D738–743, 2009.

[31] UCSC Genome Bioinformatics. `http://genome.ucsc.edu/`.

[32] Pfam: Home Page. `http://pfam.sanger.ac.uk/`.

[33] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996-2010. `http://www.repeatmasker.org`.