Current Trends in Pseudogene Detection and Characterization

Eric Christian Rouchka*,† and I. Elizabeth Cha†

Department of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, USA

Abstract: Pseudogenes are homologous relatives of known genes that have lost their ability to function as a transcriptional unit. Three classes of pseudogenes are known to exist: duplicated pseudogenes; processed or retrotransposed pseudogenes; and unitary or disabled pseudogenes. Since pseudogenes may display a number of the characteristics of functional genes, they pose a unique set of problems for *ab initio* gene prediction. The ability to detect and differentiate pseudogenes from functional genes can be a difficult task. We present a comprehensive review of current approaches for pseudogene detection, highlighting difficulties in pseudogene differentiation.

Keywords: Pseudogene, gene prediction, disabled pseudogenes, processed pseudogenes, duplicated pseudogenes.

INTRODUCTION

Bioinformatics and Computational Biology approaches have aided in the understanding of the mechanisms for gene regulation and expression. Examples of statistical models employed include expectation-maximization [1,2], Gibbs sampling [3-5] and hidden Markov model [6,7] approaches for detecting subtle regulatory regions; hidden Markov models for predicting gene coding regions [8-11], and context-free grammars and hybrid models for predicting regulatory RNAs [12-17]. Additionally, scientists are interested in understanding functional gene evolution. One source of information is pseudogenes, which are defunct relatives of known genes. The current review is devoted to introducing the existing approaches for pseudogene differentiation in hope of further stimulating development of statistical modeling approaches for understanding gene regulation and evolution.

GENES

Genes are the basic unit for transferring hereditary information. They provide structure, function and regulation to a biological system. The genetic blueprint of an organism can be understood by studying the architecture of these genes. The Central Dogma of Molecular Biology, first described by Francis Crick in 1958 and restated in 1970 [18], states that a gene must go through several steps from a genetic DNA sequence to a fully-functional protein: transcription, pre-mRNA processing, translation and protein folding Fig. (**1**). If any of these steps fails, the sequence may be considered nonfunctional. The most commonly identified disablements are stop codons and frame shifts, which almost universally stop the translation of a functional protein product.

Pseudogenes are homologous sequences arising from currently or evolutionarily active genes that have lost their ability to function as a result of disrupted transcription or translation. They may contain stop codons, repetitive elements, have frame shifts and/or lack of transcription. However, they might retain gene-like features, such as promoters, CpG islands and splice sites. Pseudogenes are of particular interest to biologists since they can interfere with gene centric studies (such as de novo gene prediction and PCR amplification). Evolutionary biologists also have an interest in pseudogenes due to the ability to study their age and mutational tendencies.

Types of Pseudogenes

In general, there are two categories of pseudogenes: 1) processed pseudogenes that have arisen out a retrotransposable event that incorporates a processed mRNA into a genomic region and 2) unprocessed, or duplicated, pseudogenes that result from a gene duplication event. Fig. (**2**) illustrates these two types of processes leading to pseudogenes. A third class of pseudogenes, unitary pseudogenes, may also exist that result from disablements within previously functional genic regions.

Processed Pseudogenes

Processed pseudogenes, also known as retrotransposed pseudogenes, arise out of retrotransposable events from mRNA. These are the most commonly studied pseudogenes, due to their distinguishable characteristics indicating RNA processing, including lack of intronic regions, poly-A tracts at the carboxyl (3') end, and flanking repeat regions. Processed pseudogenes are often the result of a reverse-transcribed mRNA that has been re-integrated into genomic DNA. Mobilization of processed mRNAs for processed pseudogene formation is likely to be caused in part by long interspersed elements (LINEs) [19,20] and endogenous retroviral elements, such as HERV-W [21]. Processed pseudogenes are found in abundance in mammalian genomes, but are relatively rare for other genomes [22]. For instance, in human chromosome 22, 19% of the coding sequence corresponds to pseudogenes, 82% of which are processed pseudogenes [23,24]. Processed pseudogenes represent the vast majority of inactivated gene sequences found in the human genome.

There are three basic features of a processed pseudogene. They are:

*Address correspondence to this author at the University of Louisville, Speed School of Engineering, Department of Computer Engineering and Computer Science, 123 JB Speed Building, Louisville, KY 40292, USA; Tel: 502-852-1695; Fax: 502-852-4713;
E-mail: eric.rouchka@louisville.edu

†Both authors contributed equally to this work

1574-8936/09 $55.00+.00 © 2009 Bentham Science Publishers Ltd

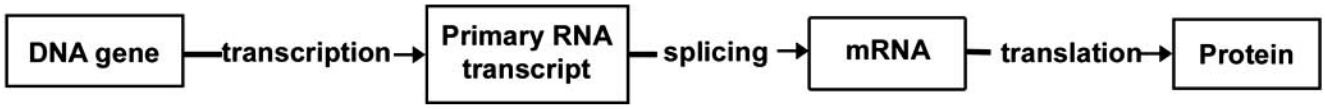

Fig. (1). Central dogma of molecular biology.

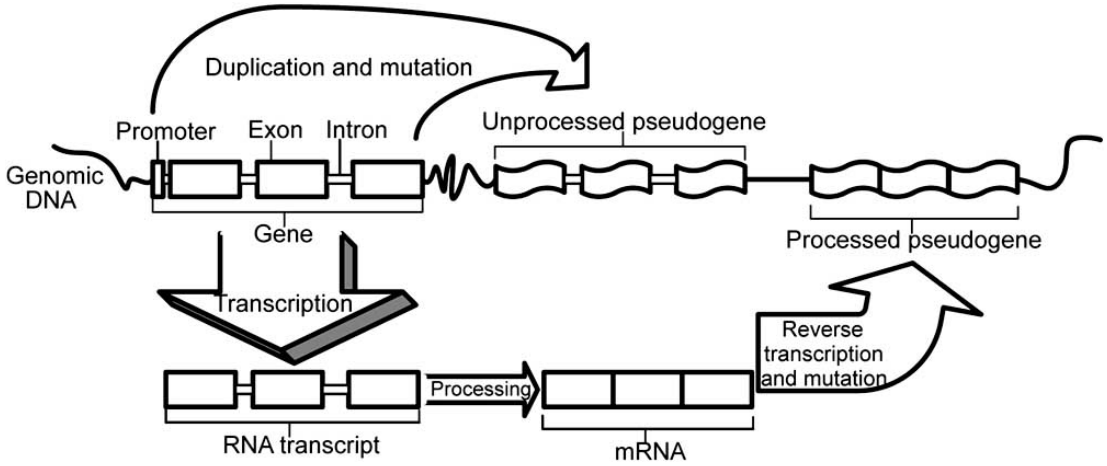

Fig. (2). Pseudogene formation.

1. Lack of non-coding intervening sequences (introns and promoters),
2. Presence of a poly-A tail at the 3' end, and
3. Homological extensions; i.e. (flanking direct repeats, which are associated with insertion sites of transposable elements) [23,25].

Often processed pseudogenes are truncated at the 5' end, due to the low processivity of reverse transcriptase [21]. In the human genome, there are a number of different estimations for the total number of processed pseudogenes. Zhang *et al.* predict ~7,800 processed pseudogenes [26] while Torrents *et al.* estimate there are ~13,800 [27]. Ohshima *et al.* report 3,664 processed pseudogenes in the human genome and estimate a total number ~7,000 processed pseudogenes, based on an estimation of ~35,000 human genes [28-30]. Palicek *et al.* [20] survey the number to be somewhere in the wide range of between 8,000 and 100,000, with the large upper bound originating from 5' truncated copies. However, this large number of pseudogenes is derived from a small percentage (10%) of actual genes [20]. Some processed pseudogenes are also disrupted by repetitive elements [31].

It has been observed that there is a correlation between the length of the coding sequence of a functional gene and the number of transcribed processed pseudogenes [32]. While this may indeed be the case, a potential issue arises with the difficulty in detecting partial processed pseudogenes.

Comparative analysis of the mouse and human genomes characterizes additional properties of processed pseudogenes [33]. They tend to correspond to functional housekeeping genes. This makes perfect sense, due to the fact that they would be expressed in germline cells, and therefore could be heritable. In addition, processed pseudogenes follow the general principle of LINEs and tend to occur in GC-poor regions of the genome.

Duplicated Pseudogenes

In the process of duplication, genetic sequences can go through various modifications, including mutation, frame shifting, insertions and deletions. Any significant modification during the translational or transcriptional stage could result in a loss of genetic function. Hence, if the duplication of a gene is incomplete, the new sequence could be a pseudogene. Those pseudogenes derived from duplication are called unprocessed, or duplicated, pseudogenes [23,34,35]. They arise when a cell is replicating its own DNA and inserts an extra copy of a gene into a genome in a new location [36].

Unprocessed pseudogenes may have introns retained (as shown in Fig. (**2**)). They can be a complete or partial copy from the parent gene. Sometimes, the unprocessed pseudogenes contain an extra copy of the gene. The unnecessary extra copy "could accumulate mutations without harming the organism" [25]. An unprocessed pseudogene could also be created secondarily from pre-existing processed pseudogenes where the parental source sequence may be intronless. In this case, the duplication occurred after transcription [23].

Unitary Pseudogenes

A third type of pseudogenes, known as disabled or unitary pseudogenes, are becoming more widely studied [37,38]. When various mutations occur, they can disrupt transcription or translation of a gene. Moreover, if these mutations become fixed in a population, it is possible the gene may become nonfunctional or deactivated in a mechanism similar to unprocessed pseudogenes. The difference between the two is that disabled pseudogenes are not duplicated before becoming disabled [39]. The difficulty in classifying unitary pseudogenes is that in actuality they could be ancient duplicated pseudogenes that have sufficiently diverged from the functional homolog in such a way that they can no longer be detected as similar. Such cases require more of an evolu-

tionary approach incorporating multiple organisms, as seen with the ψFXRβ pseudogene [40].

FUNCTIONAL OR NONFUNCTIONAL?

The traditional view of pseudogenes as genetic elements similar to functional genes without functional properties has recently come under fire [41], particularly in light of experimental evidence that there is in actuality very little "junk" DNA. The traditional view would suggest that since pseudogenes do not produce protein products, they are typically not under selective evolutionary pressure and thus evolve at rates consistent with neutral drift [22].

A challenge to this theory arises from recent evidence that a number of pseudogenes are actively transcribed. Zheng *et al.* [33,42] use tiling microarrays as evidence of transcription for as many as 20% of all known pseudogenes on human chromosome 22. Additionally, Harrison *et al.* [43] identify a set of 233 transcribed processed pseudogenes within the mouse and human genome using as a source expressed sequences in the form of mRNAs and ESTs. While none of these are translated into proteins due to sequence disablements, they are hypothesized to have roles in gene regulation by use of their sequences complementary to the homologous functional gene. Transcribed processed pseudogenes have an additional effect that they themselves can become duplicated, resulting in "duplicated-processed" pseudogenes [40,44].

However, many characterized pseudogenes have been implicated in regulation of gene expression, gene regulation, and provide a potential source of genetic diversity through recombination with functional genes or exon shuffling [20]. It is even suggested that pseudogenes should be classified as potogenes to represent their potential to become new genes [45].

Specific examples of the role of pseudogenes in expression is a retrotransposed pseudogene makorin1-p1 whose transcription plays a trans-regulatory role in the expression of its parent gene, Makorin1 [46,47]. This particular experiment has been strongly debated due to lack of reproducibility [48]. More recent studies suggest regulatory roles between pseudogenes and genes as a result of siRNA targeting. In 2008, Piehler *et al.* [49] described interactions between ABC transporter genes and pseudogenes that suggests that expression of the gene ABCC6 is regulated, in part, by transcription of the ABC6P1 pseudogene. Experimental evidence for this regulation was supported using targeted siRNA knockdown of the ABC6P1 pseudogene. Tam *et al.* [50] demonstrate the possibility that a subset of mammalian pseudogenes are responsible for generating small interfering RNAs (siRNAs) by forming dsRNA sequences with the corresponding protein-coding mRNA. These pseudogene-derived siRNAs are then in turn responsible for regulating the functional gene from which the pseudogene originates by acting to degrade the functional gene's mRNA transcripts [41,48].

A full opinion paper appeared in Trends in Genetics in 2007 in which the argument about pseudogene functionality was discussed in great detail [51]. In their discussion, Zheng and Gerstein differentiate genes and pseudogenes while further classifying pseudogenes into ghost pseudogenes that have some intermediate functionality (such as a regulatory function or transcriptional activity) and dead pseudogenes that do not have any indication of functionality, and therefore are subject to neutral drift. In order to clarify genes from pseudogenes, they offer an updated definition of pseudogenes as "genomic sequences that arise from functional genes but that cannot encode the same type of functional product (i.e. protein, tRNA or rRNA) as the original genes".

AUTOMATIC DETECTION OF PSEUDOGENES

Homology-Based Approaches

Nearly all current pseudogene detection algorithms employ an initial homology-based search to identify potential pseudogeneic regions. This approach takes a genome or chromosome of interest, many of which can be found as part of UCSC's goldenpath assemblies (genome.ucsc.edu), and compares it to a known set of protein sequences such as ENSEMBL [52], UniProt [53], or RefSeq [54]. Since this comparison involves nucleotide and amino acid sequences, the nucleotide sequence must be translated into all six possible reading frames (three on the forward strand; three on the reverse). This is typically accomplished by using a fast heuristic algorithm such as `tblastn` [55] as a first-pass.

Genomic regions not corresponding to a known gene's origin that pass a defined threshold (typically involving the percent of the gene that is covered) are then aligned to the corresponding gene sequence with finer detail using methods such as `tfastx` [56]. This will produce alignments that can be further analyzed for different types of disablements. The `tfastx` alignments are particularly useful when studying processed pseudogenes that do not contain intronic regions. For duplicated pseudogenes, an algorithm such as `sim4` [57], `spidey` [58], `est2genome` [59] or `exalign` [60] that takes into account intron/exon splice site information might prove to be useful, particularly if the reference protein product is represented as nucleotides (as is the case with EST data).

Once the genome-to-protein alignments are filtered, they are searched for disablements. Typically, this includes premature stop codons, 3' poly-A tails, 3' poly-A sites, and a non-synonymous:synonymous ratio (generally denoted as dN/dS or Ka/Ks for short) above a certain threshold, indicating neutral drift.

Harrison's Approach

Harrison *et al.* from Yale University developed a pseudogene annotation pipeline in 2001 [61]. They used *Caenorhabditis elegans* genome to test their process. First, they downloaded annotated pseudogene sequences from Sanger Centre (http://www.sanger.ac.uk) and analyzed those sequences. They found about 6% of the annotated pseudogenes are not detectable when considering simple disablement, such as a frameshift or premature stop codon. Second, they aligned protein sequences (Wormpep18, SCOP and PROTOMAP) to the genome sequences (translations in the three forward and three reverse reading frames were considered). Third, alignment results were filtered for overlapping matches in genomic sequences, if the size of the overlap is over 30 nucleotides. Fourth, to avoid over-counting for worm protein hits, the initial matches were further filtered for adjacent matches. Fifth, the potential pseudogenes were filtered for overlap with any other annotations in the Sanger

Centre files such as exons of genes, tandem or inverted repetitive and transposable elements. Sixth, they refined the data for additional repeat elements any match to proteins. Finally, they increased the threshold e-value from 0.01 to 0.001 [61]. In 2002, using the same methodology, Harrison *et al.* released their results for detecting pseudogenes in human chromosome 21 and 22. They found 77 processed pseudogenes on chromosome 21 and 112 on chromosome 22 [62].

Sakai's Approach

In 2003, Sakai *et al.* from Japan introduced a method to detect processed pseudogenes based on cDNA mapping to the human genome [63]. In the first phase, they aligned all cDNA sequences against the human genome using `blastn`. Then they extracted corresponding genomic regions for each query sequen `tblastn` ce. Next, they set up cut values of ≥ 95% identity and ≥ 80% coverage to filter those sequence hits based on `est2genome`.

The second stage of the experiment is pseudogene detection. All the candidate sequences collected from the first stage were classified into six classes according to the way in which the introns are processed before checking the 3' terminator: complete intron loss, one or more intron loss, no intron loss, no intron in the functional gene and at least one <80 bp intron.

Sakai and his team found 6,549 possible processed pseudogenes using 113,196 human cDNA with their automated pipeline including novel intron-containing pseudogenes. They believed their approach will help aid the accuracy of gene prediction [63].

PPFINDER

PPFINDER was released in 2006 to find processed pseudogenes [29]. The purpose for the program is to improve the accuracy of gene prediction methods, if processed pseudogenes can be identified. A gene model was first established in order to collect the gene information. Two methods were used in ordered to find pseudogenes: intron location and conserved synteny. In the intron location method, it started to align pre-established gene model to cDNA database using `blastn`. The best scoring hits were kept. Then, these transcripts were further aligned to genome locus. If the transcript hits more than one location in the genome, the transcript will be marked as pseudogene-derived for further testing. The segments of predicted gene will only hit themselves in the genome, if they are not pseudogene-derived. But if they are pseudogene-derived, the best hit score will be with the parental gene location. The other hit locations in the genome can be potential pseudogenes.

In the synteny method, the candidate pseudogene-derived sequences were split into exons and translated into amino acids. Then these translated sequences were aligned to the protein database using `blastp`. If the sequence identity of the hit is ≥ 65% and the length of the hit is at least 10 amino acids, these sequences were examined if their hit location is the same as the gene model. If the answer is no, the information of conserved syntenic region will be recorded. Once again, they aligned the original created gene model to the recorded information to find orthologous region. If the orthologous regions were found, the gene model is said to contain a pseudogene.

However, if it is single-exon gene, without locating introns, the sequence might be counted as a repeat and would be failed to identify as pseudogene-derived sequence [29].

Pseudogene Finder (PSF)

In 2006, another pseudogene prediction pipeline, PSF, was proposed in order to detect pseudogenes within 44 ENCODE sequences [33,64]. PSF begins by searching for protein alignments not corresponding to the gene location as determined by the program `Prot-map`. Potential pseudogenes are filtered from this resulting dataset by looking at alignments passing a threshold based on the number of aligned and non-aligned amino acids, the number of nucleotides within an exon, and the number of aligned amino acids and nucleotides outside of the reference gene sequence. There are further filtered to ensure that at least one disruption event occurs (frameshift, internal stop codon, poly-A site and/or poly-A tract at 3' end, lack of introns for 95% of the sequence; Ka/Ks ratio > 0.5). Based on these criteria, PSF was able to detect 81% of annotated pseudogenes for the corresponding ENCODE regions.

PseudoPipe

In 2006, a homology-based automated pseudogene identification pipeline known as PseudoPipe was introduced [65] which builds upon a previously discussed pipeline [26,30]. The identification process begins by aligning protein sequences to genomic sequences using `tblastn`. Sequences overlapping with the genes annotated in the ENSEMBL database (http://www.ensembl.org/) are then removed. The remaining sequences are further aligned using the Smith-Waterman approach of `tfasty` [26,30,56] to optimize the alignment to merge and extend candidate sequences. Fourth, false positive repeats, low complexity sequences and potential functional gene candidates were removed. The pseudogene candidate is considered a false-positive if its BLAST e-value is ≤ 40%. Finally, sequences are separated into different classes - retrotransposed pseudogenes, duplicated pseudogenes (or unprocessed pseudogenes in our definition), or pseudogeneic fragments based upon exon structure [26,30,32,65].

Pseudogenes Analysis in the ENCODE Project

The ENCylopedia of DNA Elements (ENCODE) project was initiated in September 2003. It brought several dozen laboratories together identify the functional elements in the human genome. Two components - a pilot phase and a technology development phase - were proposed when the project started.

The aim of the pilot phase was to evaluate a variety of different methods for use in later stages. A portion of the genome equal to about 1% (30MB) was carefully chosen and analyzed using new and existing methods. Transcripts and gene models, protein-binding sites, epigenetic marks, promoters and enhancers, and DNA replication regions were the fundamental items involved in the research. The conclusion of this pilot project were published in June 2007 [66-75].

The technology development phase promoted several new technologies and protocols to generate high throughput data on functional elements for use in the ENCODE project. A third phase the production phase began in September 2007. A Data Coordination Center at the University of California at Santa Cruz (UCSC) was created to track, store and display the ENCODE data. In addition, a Data Analysis Center was established. All data generated by the ENCODE participants is released by the Data Coordination Center [70].

In the ENCODE project, one group focused on methods for detecting pseudogenes. They established methods to decide consensus annotation, analyze transcription and study evolution. The first goal was to "obtain an accurate list of pseudogenes in order to facilitate the creation of a comprehensive catalog of structural and functional elements in the ENCODE regions" [66,68]. They evaluated five developed methods: GIS-PET [76], HAVANA [77], PseudoPipe [65,67], pseudoFinder (unpublished), and retroFinder [78,79].

The pseudogene research team found that all five methods detect pseudogenes based on sequence similarity to "at least one entry in a collection of query sequences representing known human genes (referred to as the parent genes)." The main two differences among these five methods are in (1) type of queries, either nucleotides or proteins, used to search for pseudogenes and (2) strategies used to assess a sequence's coding potential and methods for distinguishing processed from unprocessed pseudogenes [68].

The second task was to establish a consensus procedure including manual curation in order to get the correct and reliable list of pseudogenes. The research team only considered those pseudogenes that match known proteins from the UniProt database [80].

After evaluating the five computational methods, they characterized the sequences they found and sequenced these pseudogene transcripts. They realized that in the ENCODE gene-rich regions, more pseudogenes were found. In addition, most sequences are decayed gene copies that contain frame shift mutations [68].

Classifier of Olfactory Receptor Pseudogenes (CORP)

In 2006, Menashe *et al.* [81] described a probabilistic classifier for distinguishing between functional and nonfunctional olfactory receptors (OR) from the mammalian gene family, which contains between 900-1400 genes (half of which are thought to be pseudogenes). Their approach, which they labeled Classifier of Olfactory Receptor Pseudogenes, or CORP, is a probabilistic model based upon conservation of critical residues which are determined by multiple sequence alignment of the OR family from different species. Individual sequences are then compared to the "crucial consensus" of 65 detected critical bases for conservation measures. This approach works well for identifying functional genes (95%) but suffers by missing 35% of nonfunctional pseudogenes. This classifier may prove useful when dealing with known gene families, but it requires prior knowledge in order to create an appropriate training set. Therefore, it cannot be used for *de novo* pseudogene detection.

PSEUDOGENE DATABASES

HOPPSIGEN

HOPPSIGEN [82] is a database of homologous processed pseudogenes shared between the mouse and human genomes that contains information pertaining to the genomic location, potential function, and gene structure of processed pseudogenes.

PseudoGeneQuest

PseudoGeneQuest [83] is a web-based system for identifying novel human pseudogenes based on a user provided protein sequence, which is often derived from EST data. The implementation of PseudoGeneQuest begins by comparing the query sequence against the human genome using `tblastn`. Hits with an e-value $< 10^{-4}$ are then compared to known genes and pseudogenes from pseudogene.org [84]. Using criteria based on repetitive element inclusion, sequence coverage, amino acid identity, premature stop codons and frame shifts, genome regions are classified into nine categories: 1) already known genes; 2) known pseudogenes; 3) real gene or exon; 4) pseudogene fragment; 5) pseudogene; 6) miscellaneous; 7) putative new gene; 8) duplicated pseudogene; 9) interrupted processed pseudogene.

Pseudogene.Org

Pseudogene.org [84] is a MySQL-based repository for pseudogenes identified using various methods, including over 100,000 pseudogenes from 75 genomes. The pseudogene.org database has been created with features to: 1) allow for the integration of pseudogene information identified from different sources; 2) allow for flexible search utilities; 3) store predefined pseudogene sets of interest; 4) interact in a robust fashion with existing pseudogene databases; 5) reconstruct historical pseudogene data based on temporal information; and 6) simply database accessibility using Perl libraries. Pseudogenes within Pseudogene.org are classified into the four categories of 1) processed pseudogenes; 2) duplicated pseudogenes; 3) other non-processed pseudogenes (such as unitary pseudogenes); and 4) unclassified (due to signal degradation or structural ambiguity).

Pseudogene.org classifies pseudogene detection methods according to an entity attribute value (EAV) that includes the Ka/Ks ratio; CpG content; distance from the query protein; proximity to CpG islands; and relevant PCR tiling microarray results.

University of Iowa Pseudogene Resource

The University of Iowa's UI Pseudogenes website (http://genome.uiowa.edu/pseudogenes/) contains 14,476 pseudogenes identified in the human genome. Included in this database are 7,146 processed pseudogenes, 3,426 non-processed pseudogenes and 3,904 pseudogenes with ambiguous classification. Methods for classifying these pseudogenes is discussed in the manuscript by Bischof *et al.* [85]

DISCUSSION

Inconsistently Detected Pseudogenes

One of the main difficulties with computational detection of pseudogenes is that many of the current methods in place only share a small percentage of detected pseudogenes.

Karro *et al.* [84] illustrate this by showing a Venn diagram of pseudogenes identified by PseudoPipe, Hoppsigen, and Torrents *et al.* [27]. This figure shows that there are only 903 pseudogenes detected by all three methods. In fact, the majority of the pseudogenes detected by these methods are uniquely identified based on the computational approach. Torrents detects 11,585 unique pseudogenes, PseudoPipe detects 8965 unique pseudogenes, and Hoppsigen detects 2149 unique pseudogenes.

Performance analysis of pseudogene prediction approaches is an area in which great improvement is needed. The overriding difficulty is the lack of known true positive pseudogenes. Currently, most pseudogenes are labeled as such due to their detection using a computational approach. However, use of these as a training set causes validation issues, since in actuality it will only measure whether or not two separate approaches can detect the same set types of pseudogenes. What is desperately needed for analyzing the performance of pseudogenes is a "gold standard" test set similar to those found for use in analyzing gene prediction programs [86]. Such a set would include examples of processed, duplicated, and unitary pseudogenes. Once such a set is in place, each approach can be evaluated using sound statistical tests such as the independent dataset, subsampling, and jackknife tests [87-89]. As it currently stands, each of these approaches should be used as first-pass filters whereby the results are evaluated as potential pseudogenes for further study. The inconsistent results of the approaches discussed in this review illustrate the major difficulty of validation.

Difficult Cases

Pavlicek *et al.* [21] show that only 28% of HERV-W derived processed pseudogenes are full length while in silico methods report around 95% of all processed pseudogenes as full length [90]. This suggests that current computational pseudogene detection algorithms may be severely underestimating the total number of processed pseudogenes. It is suggested that there are likely to be three times as many processed pseudogenes than are currently characterized [21].

Pseudogenes that truly become non-functional present a particular problem in pseudogene detection. Since these are susceptible to neutral drift, they will over a period of time lose their ability to be recognized by computational methods, particularly in light of differing rates of deletions of genomic regions [91,92]. One method for overcoming this issue is to detect partially conserved protein motifs in intergenic regions called pseudomotifs [90].

Processed pseudogene formation is not an ancient event, but one that seems to be actively occurring within mammalian genomes [19,93] with the help of LINEs. This presents a challenge in their detection and age determination while at the same time providing for genetic diversity by allowing for insertion of elements that may retain a transcriptional function.

Duplication of gene sequences causes a particular problem. Just because two sequences have sufficiently diverged, it does not necessarily mean that one is a pseudogene. In fact, the delineation between paralogous genes and duplicated pseudogenes can be extremely difficult. Efforts to detect and annotate paralogous genes, such as with ParaDB [94] and u-Genome [95], have begun to differentiate between paralogs and duplicated pseudogenes. However, these datasets are incomplete, and still suffer from many of the same issues of being derived from sequence level homology which begs the question as to whether these are indeed paralogs or are pseudogenes. In some cases, genes that appear to be functional might in reality be duplicated pseudogenes that have been disabled due to disruptions in the promoter or splice signaling mechanisms [61,91,92].

Nearly all computational detection methods currently rely on comparing a genomic region to a known protein sequence in order to find novel pseudogenes. However, this approach does not allow for the detection of unitary pseudogenes where a reference gene is absent. In the same light, limitations of sequence alignment approaches make it hard to detect ancient and decayed pseudogenes. Single exon genes present a unique matching problem, since any associated pseudogene is likely to be classified as a processed pseudogene due to the lack of introns.

Future Directions

As more genomes of higher level organisms become sequenced, the ability to do comparative genomics studies of pseudogenes will become more plausible. This is particularly important in cases where processed pseudogenes are ancient and in cases where unitary pseudogenes have formed. Incorporation of comparative data will likely increase both the sensitivity and specificity of pseudogene detection approaches, albeit at an increased price in computation.

The Ka/Ks ratio is commonly used to determine whether or not a particular region has diverged, by accumulating a number of nonsynonymous mutations. In some cases where sequences have diverged, this might lead to labeling true genes as pseudogenes. A more accurate measure might consider looking at similar-synonymous mutations where a codon encoding for one amino acid might mutate to code for a separate amino acid with similar properties.

While transposable elements have been traditionally thought of as contributing to "junk DNA", they offer a mechanism for gene diversity and transformation. Evidence of fewer processed pseudogenes in fruit flies [96], yeast, and worms [61] as opposed to mammalian genomes may point in part to the potential for gene duplication events in higher organisms. As more knowledge is obtained regarding biological mechanisms at the genome level (such as the discovery of miRNA), it is possible that truly nonfunctional regions of the genome will be discovered to be rare events. As a result, the fine line between nonfunctional pseudogenes and functional components in the genome will continue to be of interest for years to come.

ACKNOWLEDGEMENTS

Support provided by NIH NCRR grant P20RR16481 and NIH NIEHS grant P30ES014443. Its contents are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

REFERENCES

[1] Stormo GD, Hartzell GW, III. Identifying protein-binding sites from unaligned DNA fragments**.** *Proc Natl Acad Sci USA* **1989**; 86: 1183-7.

[2] Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **1990**; 7: 41-51.

[3] Lawrence CE, Altschul SF, Boguski MS, *et al*. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**; 262: 208-14.

[4] Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* **1995**; 4: 1618-32.

[5] Thompson W, Rouchka EC, Lawrence CE. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **2003**; 31: 3580-5.

[6] Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN. Promoter region-based classification of genes. *Pac. Symp Biocomput* **2001**; 151-63.

[7] Grundy WN, Bailey TL, Elkan CP, Baker ME. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci* **1997**; 13: 397-406.

[8] Borodovsky M, Mills R, Besemer J, Lomsadze A. Prokaryotic gene prediction using GeneMark and GeneMark.hmm. *Curr Protoc Bioinformatics* **2003**; Chapter 4: Unit4.

[9] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **1997**; 268: 78-94.

[10] Korf I. Gene finding in novel genomes. *BMC Bioinformatics* **2004**; 5: 59.

[11] Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Comput Appl Biosci* **2001**; 17 (Suppl 1): S140-S148.

[12] Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci USA* **2008**; 105: 14885-90.

[13] Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks. *Comput Appl Biosci* **2008**; 24: i50-i58.

[14] Ritchie W, Theodule FX, Gautheret D. Mireval: a web tool for simple microRNA prediction in genome sequences. *Comput Appl Biosci* **2008**; 24: 1394-6.

[15] Griffiths-Jones S, Saini HK, van DS, Enright AJ. miRBase: tools for microRNA genomics2. *Nucleic Acids Res* **2008**; 36: D154-D158.

[16] Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK. Naive Bayes for microRNA target predictions--machine learning for microRNA targets. *Comput Appl Biosci* **2007**; 23: 2987-92.

[17] Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. *Cell* **2003**; 115: 787-98.

[18] Crick F. Central dogma of molecular biology. *Nature* **1970**; 227: 561-3.

[19] Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **2000**; 24: 363-7.

[20] Pavlicek A, Gentles AJ, Paces J, Paces V, Jurka J. Retroposition of processed pseudogenes: the impact of RNA stability and translational control. *Trends Genet* **2006**; 22: 69-73.

[21] Pavlicek A, Paces J, Zika R, Hejnar J. Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. *Gene* **2002**; 300: 189-94.

[22] Friedberg F, Rhoads AR. Calculation and verification of the ages of retroprocessed pseudogenes. *Mol Phylogenet Evol* **2000**; 16: 127-30.

[23] Cooper DN. Pseudogenes and their formation. In Human Gene Evolution, BIOS Scientific Publishers, Oxford, UK **1999**; 265-85.

[24] Dunham I, Shimizu N, Roe BA, *et al*. The DNA sequence of human chromosome 22. *Nature* **1999**; 402: 489-95.

[25] Gibson LJ. Pseudogenes and Origins. *Origins* **1994**; 21: 91-108.

[26] Zhang Z, Harrison PM, Liu Y, Gerstein M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res* **2003**; 13: 2541-58.

[27] Torrents D, Suyama M, Zdobnov E, Bork P. A Genome-Wide Survey of Human Pseudogenes. *Genome Res* **2003**; 13: 2559-67.

[28] Ohshima K, Hattori M, Yada T, *et al*. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* **2003**; 4: R74.

[29] van Baren MJ, Brent MR. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome Res* **2006**; 16: 678-85.

[30] Zhang Z, Gerstein M. Large-scale Analysis of Pseudogenes in the Human Genome. *Curr Opin Genet Dev* **2004**; 14: 328-35.

[31] Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **2002**; 12: 1466-82.

[32] Zhang Z, Gerstein M. Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome. *Genomics* **2003**; 81: 468-80.

[33] Zhang Z, Carriero N, Gerstein M. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet* **2004**; 20: 62-7.

[34] Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. *FEBS Lett* **2000**; 468: 109-14.

[35] Vanin EF. Processed Pseudogenes: Characteristics and Evolution. *Ann Rev Gene* **1985**; 1985: 253-72.

[36] Gerstein M, Zheng D. The real life of pseudogenes. *Sci Am* **2006**; 295: 48-55.

[37] Nishikimi M, Kawai T, Yagi K. Guinea pigs possess a highly mutated gene for L-gulono-gamma-lactone oxidase, the key enzyme for L-ascorbic acid biosynthesis missing in this species. *J Biol Chem* **1992**; 267: 21967-72.

[38] Nishikimi M, Fukuyama R, Minoshima S, Shimizu N, Yagi K. Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *J Biol Chem* **1994**; 269: 13685-8.

[39] Xue Y, Daly A, Yngvadottir B, *et al*. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *Am J Hum Genet* **2006**; 78: 659-70.

[40] Zhang ZD, Cayting P, Weinstock G, Gerstein M. Analysis of nuclear receptor pseudogenes in vertebrates: how the silent tell their stories. *Mol Biol Evol* **2008**; 25: 131-43.

[41] Sasidharan R, Gerstein M. Genomics: protein fossils live on as RNA. *Nature* **2008**; 453: 729-31.

[42] Zheng D, Zhang Z, Harrison PM, *et al*. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol* **2005**; 349: 27-45.

[43] Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* **2005**; 33: 2374-83.

[44] Zhang Z, Gerstein M. The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene* **2003**; 312: 61-72.: 61-72.

[45] Balakirev ES, Ayala FJ. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* **2003**; 37: 123-51.

[46] Hirotsune S. An expressed pseudogene regulates mRNA stability of its homologous coding gene. *PNE* **2003**; 48: 1908-12.

[47] Hirotsune S, Yoshida N, Chen A, *et al*. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* **2003**; 423: 91-6.

[48] Gray TA, Wilson A, Fortin PJ, Nicholls RD. The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proc Natl Acad Sci USA* **2006**; 103: 12039-44.

[49] Piehler AP, Hellum M, Wenzel JJ, *et al*. The human ABC transporter pseudogene family: Evidence for transcription and gene-pseudogene interference. *BMC Genomics* **2008**; 9: 165.

[50] Tam OH, Aravin AA, Stein P, *et al*. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **2008**; 453: 534-8.

[51] Zheng D, Gerstein MB. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* **2007**; 23: 219-24.

[52] Flicek P, Aken BL, Beal K, *et al*. Ensembl 2008. *Nucleic Acids Res* **2008**; 36: D707-D714.

[53] The universal protein resource (UniProt). *Nucleic Acids Res* **2008**; 36: D190-D195.

[54] Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2008**; 36: D13-D21.

[55] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool3. *J Mol Biol* **1990**; 215: 403-10.

[56] Pearson W. Finding protein and nucleotide similarities with FASTA**.** *Curr Protoc Bioinformatics* **2004**; Chapter 3: Unit3.
[57] Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence**.** *Genome Res* **1998**; 8: 967-74.
[58] Wheelan SJ, Church DM, Ostell JM. Spidey: a tool for mRNA-to-genomic alignments**.** *Genome Res* **2001**; 11: 1952-7.
[59] Churbanov A, Pauley M, Quest D, Ali H. A method of precise mRNA/DNA homology-based gene structure prediction**.** *BMC Bioinformatics* **2005**; 6: 261.
[60] Zhang M, Gish W. Improved spliced alignment from an information theoretic approach**.** *Comput Appl Biosci* **2006**; 22: 13-20.
[61] Harrison PM, Echols N, Gerstein MB. Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome**.** *Nucleic Acids Res* **2001**; 29: 818-30.
[62] Harrison PM, Hegyi H, Balasubramanian S, *et al*. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22**.** *Genome Res* **2002**; 12: 272-80.
[63] Sakai H, Koyanagi KO, Itoh T, Imanishi T, Gojobori T. Detection of Processed Pseudogenes Based on cDNA Mapping to the Human Genome**.** *Genome Informat* **2003**; 14: 452-3.
[64] Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters**.** *Genome Biol* **2006**; 7 (Suppl 1): S10-S12.
[65] Zhang Z, Carriero N, Zheng D, *et al*. PseudoPipe: an automated pseudogene identification pipeline**.** *Comput Appl Biosci* **2006**; 22: 1437-9.
[66] The ENCODE (ENCyclopedia Of DNA Elements) Project**.** *Science* **2004**; 306: 636-40.
[67] Zheng D, Gerstein MB. A computational approach for identifying pseudogenes in the ENCODE regions**.** *Genome Biol* **2006**; 7 (Suppl 1): S13-0.
[68] Zheng D, Frankish A, Baertsch R, *et al*. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Res* **2007**; 17: 839-51.
[69] Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**.** *Nature* **2007**; 447: 799-816.
[70] The ENCODE Project: ENCyclopedia Of DNA Elements**.** National Human Genome Research Institute - National Institutes of Health, 2008.
[71] Denoeud F, Kapranov P, Ucla C, *et al*. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions**.** *Genome Res* **2007**; 17: 746-59.
[72] Gerstein MB, Bruce C, Rozowsky JS, *et al*. What is a gene, post-ENCODE? History and updated definition**.** *Genome Res* **2007**; 17: 669-81.
[73] Rozowsky JS, Newburger D, Sayward F, *et al*. The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci. *Genome Res* **2007**; 17: 732-45.
[74] Thomas DJ, Rosenbloom KR, Clawson H, *et al*. The ENCODE Project at UC Santa Cruz**.** *Nucleic Acids Res* **2007**; 35: D663-D667.
[75] Zhang ZD, Paccanaro A, Fu Y, *et al*. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions**.** *Genome Res* **2007**; 17: 787-97.
[76] Chiu KP, Ariyaratne P, Xu H, *et al*. Pathway aberrations of murine melanoma cells observed in Paired-End diTag transcriptomes**.** *BMC Cancer* **2007**; 7: 109.
[77] Harrow J, Denoeud F, Frankish A, *et al*. GENCODE: producing a reference annotation for ENCODE**.** *Genome Biol.* **2006**; 7 (Suppl 1): S4-S9.
[78] Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes**.** *Proc Natl Acad Sci USA* **2003**; 100: 11484-9.
[79] Schwartz S, Kent WJ, Smit A, *et al*. Human-mouse alignments with BLASTZ**.** *Genome Res* **2003**; 13: 103-7.
[80] Bairoch A, Apweiler R, Wu CH, *et al*. The Universal Protein Resource (UniProt)**.** *Nucleic Acids Res* **2005**; 33: D154-D159.
[81] Menashe I, Aloni R, Lancet D. A probabilistic classifier for olfactory receptor pseudogenes**.** *BMC Bioinformatics* **2006**; 7: 393.
[82] Khelifi A, Duret L, Mouchiroud D. HOPPSIGEN: a database of human and mouse processed pseudogenes**.** *Nucleic Acids Res* **2005**; 33: D59-D66.
[83] Ortutay C, Vihinen M. PseudoGeneQuest - service for identification of different pseudogene types in the human genome**.** *BMC Bioinformatics* **2008**; 9: 299.
[84] Karro JE, Yan Y, Zheng D, *et al*. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation**.** *Nucleic Acids Res* **2007**; 35: D55-D60.
[85] Bischof JM, Chiang AP, Scheetz TE, *et al*. Genome-wide identification of pseudogenes capable of disease-causing gene conversion**.** *Hum Mutat* **2006**; 27: 545-52.
[86] Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences**.** *Genome Res* **2000**; 10: 1631-42.
[87] Chou KC, Zhang CT. Prediction of protein structural classes 224**.** *Crit Rev. Biochem. Mol Biol* **1995**; 30: 275-349.
[88] Chou KC, Shen HB. Recent progress in protein subcellular location prediction**.** *Anal Biochem* **2007**; 370: 1-16.
[89] Chou KC, Shen HB. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat Protoc* **2008**; 3: 153-62.
[90] Zhang ZL, Harrison PM, Gerstein M. Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes**.** *J Mol Biol* **2002**; 323: 811-22.
[91] Petrov DA, Lozovskaya ER, Hartl DL. High intrinsic rate of DNA loss in Drosophila**.** *Nature* **1996**; 384: 346-9.
[92] Petrov DA, Hartl DL. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups**.** *Mol Biol Evol* **1998**; 15: 293-302.
[93] Maestre J, Tchenio T, Dhellin O, Heidmann T. mRNA retroposition in human cells: processed pseudogene formation**.** *EMBO J* **1995**; 14: 6333-8.
[94] Leveugle M, Prat K, Perrier N, Birnbaum D, Coulier F. ParaDB: a tool for paralogy mapping in vertebrate genomes**.** *Nucleic Acids Res* **2003**; 31: 63-7.
[95] Sakharkar KR, Chaturvedi I, Chow VT, *et al*. u-Genome: a database on genome design in unicellular genomes**.** *In Silico Biol* **2005**; 5: 611-5.
[96] Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. Identification of pseudogenes in the Drosophila melanogaster genome**.** *Nucleic Acids Res.* **2003**; 31: 1033-7.

Received: September 24, 2008 Revised: October 08, 2008 Accepted: November 10, 2008