BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i94–i100
DOI: 10.1093/bioinformatics/bth942

Improved techniques for the identification of pseudogenes

L. Coin* and R. Durbin

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Pseudogenes are the remnants of genomic sequences of genes which are no longer functional. They are frequent in most eukaryotic genomes, and an important resource for comparative genomics. However, pseudogenes are often mis-annotated as functional genes in sequence databases. Current methods for identifying pseudogenes include methods which rely on the presence of stop codons and frameshifts, as well as methods based on the ratio of non-silent to silent nucleotide substitution rates (dN/dS). A recent survey concluded that 50% of human pseudogenes have no detectable truncation in their pseudo-coding regions, indicating that the former methods lack sensitivity. The latter methods have been used to find sets of genes enriched for pseudogenes, but are not specific enough to accurately separate pseudogenes from expressed genes.

Results: We introduce a program called pseudogene inference from loss of constraint (PSILC) which incorporates novel methods for separating pseudogenes from functional genes. The methods calculate the log-odds score that evolution along the final branch of the gene tree to the query gene has been according to the following constraints:

- A neutral nucleotide model compared to a Pfam domain encoding model ($\mathrm{PSILC}_{\mathrm{nuc/dom}}$);
- A protein coding model compared to a Pfam domain encoding model ($\mathrm{PSILC}_{\mathrm{prot/dom}}$).

Using the manual annotation of human chromosome 6, we show that both these methods result in a more accurate classification of pseudogenes than dN/dS when a Pfam domain alignment is available.

Availability: PSILC is available from http://www.sanger.ac.uk/Software/PSILC

Contact: lc1@sanger.ac.uk

INTRODUCTION

Pseudogenes have been defined as sequences of genomic DNA which are originally derived from functional genes but are no longer translated into functional protein products. Pseudogenes are thought to have arisen by two distinct processes. Unprocessed pseudogenes are believed to have arose from genome duplication, with a subsequent loss of function of one copy due to the accumulation of disabling mutations in the coding or regulatory sequence. Processed pseudogenes lack introns, and are thought to have arisen by reverse transcription of processed mRNA, followed by integration into the genome. Pseudogenes are increasingly thought to play an important biological role, particularly in eukaryotic genomes (Balakirev and Ayala, 2003). Duplications are believed to be a major source for the formation of new gene expression patterns and functions (Prince and Pickett, 2002). It had been assumed that due to non-functionality a pseudogene will rapidly degenerate and become indistinguishable from surrounding genomic sequence. This process has been observed in prokaryotic genomes (Andersson and Andersson, 2001). However, eukaryotic genomes contain many pseudogenes which have avoided full degeneration, and there appears to be less pressure to delete pseudogenes in eukaryotes than prokaryotes (Mighell *et al*., 2000; Harrison and Gerstein, 2002). A regulatory role for a human pseudogene has been observed experimentally (Hirotsune *et al.*, 2003).

Pseudogenes are often mis-annotated as functional genes in sequence databases (Mounsey *et al*., 2002). Two recent surveys (Torrents *et al*., 2003; Harrison *et al*., 2002) both estimate ≈20 000 human pseudogenes. Sequence based methods for identifying pseudogenes include methods which rely on the presence of truncations and methods which are based on estimating the ratio of the rates of substitution at synonymous sites to the rate of substitution at non-synonymous sites. Torrents *et al*. (2003) concluded that half of human pseudogenes have no detectable frameshifts or internal stop codons. There are many ways to estimate the rates of synonymous and non-synonymous substitution (see Bierne and Eyre-Walker 2003 for a review). In this paper, we test the method in Goldman and Yang (1994), which is commonly used, and was used in the survey from Torrents *et al*. (2003).

Here we take a novel approach to pseudogene detection, looking at the pattern of substitution in conserved protein domains. Protein domains are the structural, functional and

*To whom correspondence should be addressed.

evolutionary units of proteins. Profile hidden Markov Models (profile HMMs) are currently the most sensitive tools for identifying protein domains (Park *et al.*, 1998). The sensitivity of profile HMMs can be improved by considering the context of surrounding domains on multi-domain proteins (Coin *et al.*, 2003); by improving smoothing techniques for emission states; by considering the tree structure of the seed alignment (Qian and Goldstein, 2003) and by iterating model development. This paper uses protein domain profile HMMs from Pfam (Bateman *et al.*, 2004) to measure divergence of pseudogene members of a consensus away from that consensus, rather than testing only for membership.

ALGORITHM

Pseudogene inference from loss of constraint (PSILC) takes an alignment A, an unrooted tree T and a profile HMM D representing a Pfam domain which is aligned to A. The output is a score for each leaf-node n representing our belief that the node is a pseudogene. This score is calculated by assuming a null model of protein domain evolution on the tree, and testing the hypothesis that evolution along the final branch to the query node evolved by an alternative drift model. The score is the log-odds ratio of the probabilities that evolution along the final branch to the node has evolved under the constraint of

(1) Neutral (non-coding) DNA compared to the null Pfam domain model,

(2) Protein coding compared to the null Pfam domain model.

If a node is a pseudogene, then it is ‘released’ from the Pfam domain constraint and so we expect both scores to be higher on pseudogenes than on protein-coding genes.

We denote the row corresponding to leaf-node n by $x_{n.}$; the i-th column by $x_{.i}$ and the j-th match column of the profile HMM by m_j. We also denote by p_n the parent node of n and by b_n the branch from p_n to n. See Figure 1 for a diagramatic representation of these elements. We can calculate the probability that the alignment has evolved along the tree according to any combination of the following constraints on each branch b in the tree:

(1) neutral DNA, $\mathbf{P}_{\rm nuc}(b)$;

(2) protein coding, $\mathbf{P}_{\rm prot}(b)$;

(3) domain encoding, $\mathbf{P}_{\rm dom}(b)$.

We need to calculate probabilities under following combinations of constraints,

(1) $C_{\rm nuc} = \{\mathbf{P}_{\rm nuc}(b_n), \mathbf{P}_{\rm dom}(T\backslash b_n)\}$: neutral DNA on b_n otherwise domain encoding;

(2) $C_{\rm prot} = \{\mathbf{P}_{\rm prot}(b_n), \mathbf{P}_{\rm dom}(T\backslash b_n)\}$: protein coding on b_n otherwise domain encoding;

(3) $C_{\rm dom} = \mathbf{P}_{\rm dom}(T) = \{\mathbf{P}_{\rm dom}(b_n), \mathbf{P}_{\rm dom}(T\backslash b_n)\}$: domain encoding on all T, including b_n.

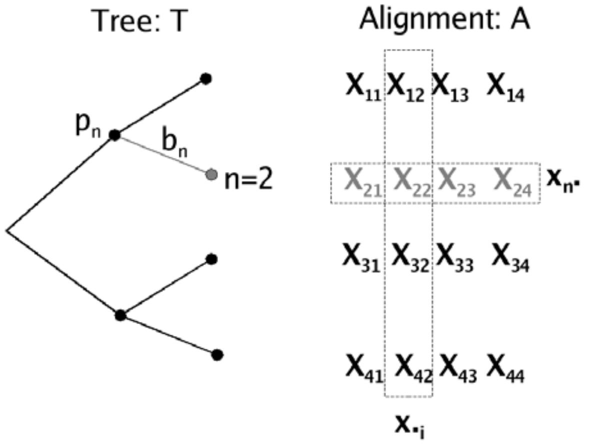
Tree: T
Alignment: A
p_n
b_n
n=2
X_{11} X_{12} X_{13} X_{14}
X_{21} X_{22} X_{23} X_{24}
$x_{n\cdot}$
X_{31} X_{32} X_{33} X_{34}
X_{41} X_{42} X_{43} X_{44}
$x_{\cdot i}$

Fig. 1. Diagram of tree T and alignment A. The tree T consists of all the branches and is unrooted (drawn here as rooted for diagramatic purposes). The node under consideration in this diagram is node 2. The final branch to this node b_n and the parent of this node p_n are labelled. We denote by $T\backslash b_n$ the tree consisting of all the black branches (i.e. excluding b_n). The alignment A corresponds to the tree T and consists of all rows. The row of the alignment corresponding to node n is labelled $x_{n.}$. We denote by $A\backslash x_{n.}$ the alignment consisting of all the black rows, which corresponds to the tree $T\backslash b_n$. We denote a column of the alignment by $x_{.i}$. We will also denote by $x_{.i}\backslash x_{ni}$ the column excluding x_{ni}.

The PSILC scores are defined as

$$\mathrm{PSILC_{nuc/dom}(n) = log\left[\frac{P(x_{n.}|A\backslash x_{n.}, T, C_{nuc})}{P(x_{n.}|A\backslash x_{n.}, T, C_{dom})}\right]}. \quad (1)$$

$$\mathrm{PSILC_{prot/dom}(n) = log\left[\frac{P(x_{n.}|A\backslash x_{n.}, T, C_{prot})}{P(x_{n.}|A\backslash x_{n.}, T, C_{dom})}\right]}. \quad (2)$$

We proceed in a manner similar to Felsenstein (1981). We assume that each x_{ni} in the row x_n is conditionally independent (given the other rows of the alignment $A\backslash x_n$, the tree T and the constraint C_k) of the other entries $x_{ni'}, i' \neq i$, so that

$$\begin{aligned} P(x_{n.}|A\backslash x_{n.}, T, C_k) &= \prod_i P(x_{ni}|A\backslash x_{n.}, T, C_k) \\ &= \prod_i P(x_{ni}|x_{.i}\backslash x_{ni}, T, C_k), \quad (3) \\ &= \prod_i \sum_{x_{p_n i} \in \Omega} \\ &\quad \times P(x_{p_n i}|x_{.i}\backslash x_{ni}, T\backslash b_n, \mathbf{P}_{\rm dom}(T\backslash b_n)) \\ &\quad \times P(x_{ni}|x_{p_n i}, b_n, \mathbf{P}_k(b_n)), \quad (4) \end{aligned}$$

where Ω denotes the alphabet of possible residues/bases at the parent node p_n. For Equation (3) we have also assumed that each each x_{ni} is conditionally independent (given $x_{.i}\backslash x_{ni}$, the tree T and the constraint C_k) of all the other columns in the alignment. For Equation (4) we have used the tree property that a leaf-node is conditionally independent of all other nodes in the tree given its parent. In the case $k = \mathrm{dom}$ we replace each $\mathbf{P}_k$ with $\mathbf{P}_{j_i}$ where j_i is the profile HMM state aligned to the column $x_{.i}$, or P_{prot} if this column aligns to an insert state. The constraint P_{j_i} represents the constraint imposed by the j_i^{th} match state of the profile HMM.

We see from Equation (4) that our algorithm comprises two steps.

(1) Calculate the frequency distribution (over residues/bases) at the parent node given the evolutionary constraints on all branches excluding the branch to the query node.

(2) For each possible residue/base at the parent node, calculate the transition probability to the child node assuming the appropriate evolutionary constraints on the branch to the child node.

To accomplish the first step we construct a new tree from the initial tree by re-rooting the tree at the parent node p_n and removing the branch to the node n. This new tree is denoted $T\backslash b_n$. Using the definition of conditional probability,

$$\begin{aligned} &P(x_{p_n i} | x_{.i} \backslash x_{ni}, T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n)) \\ &\quad = P(x_{.i} \backslash x_{ni} | x_{p_n i}, T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n)) \\ &\qquad \times \frac{P(x_{p_n i} | T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n))}{P(x_{.i} \backslash x_{ni} | T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n))} \qquad (5) \\ &\quad \propto P(x_{.i} \backslash x_{ni} | x_{p_n i}, T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n)) \\ &\qquad \times P(x_{p_n i} | T\backslash b_n, \mathbf{P}_{\mathrm{dom}}(T\backslash b_n)), \qquad (6) \end{aligned}$$

because the denominator in Equation (5) is a constant which is independent of $x_{p_n i}$. We can calculate the normalization constant by summing Equation (6) over all possible bases/residues. The first term in Equation (6) is just the likelihood of the reduced alignment conditional on each possible residue/base at the root of $T\backslash b_n$, which can be calculated using the Felsenstein (1981) algorithm. The second term in Equation (6) is the prior probability at the root given the evolutionary constraints (keeping in mind that p_n is now the root of the tree $T\backslash b_n$). We use the equilibrium distribution of the rate matrix as the prior distribution at the root, which, based on the formulation described below, corresponds to the observed distribution of bases/residues in the alignment for the DNA and protein models, respectively and to the match-state emission frequency distributions for the domain model.

For the second step as well as for the first we must calculate the probability of transitioning between different bases/residues given different evolutionary constraints. We follow the standard phylogenetic formalization of evolutionary constraints (see Lio and Goldman 1998 for a review) where the transition probability at time t is calculated as the exponential of the instantaneous rate matrix $\mathbf{Q}_{ij}$. We also follow the formulation of Goldman and Whelan (2002) in order to modify the rate matrix with respect to a steady-state distribution over residues/bases, π. With a slight abuse of notation, we will write $\mathbf{P}_k(t)$ for the matrix of transition probabilities at time t under the evolutionary constraint $\mathbf{P}_k$, and $\mathbf{P}_k(x, x', t)$ for the probability of transitioning from x to x' over time t. The equations from the above papers which concern us here are

$$\mathbf{P}_k(t) = \exp(\mathbf{Q}rt), \qquad (7)$$

$$\mathbf{Q}_{ij} = \left(\frac{\pi_j}{\hat{\pi}_j}\right)^{1-f} \times \hat{\mathbf{Q}}_{ij} \times \left(\frac{\hat{\pi}_i}{\pi_i}\right)^{f}, \qquad (8)$$

where r is a rate parameter. For amino acid models, $\hat{\mathbf{Q}}$, $\hat{\pi}$ are database estimates [e.g. the WAG model (Whelan and Goldman, 2001)] and π is the steady-state frequency specific to model in question. For nucleotide models, $\hat{\mathbf{Q}}$ is a parameterized model [e.g. the HKY model (Hasegawa *et al.*, 1985)] and $\hat{\pi}$ is the uniform distribution. For both the protein $\mathbf{P}_{\mathrm{prot}}$ and DNA $\mathbf{P}_{\mathrm{nuc}}$ models, we used the observed frequencies in the alignment for π. The free parameter f corresponds to the trade-off between frequencies in the equilibrium distribution resulting from pressure to mutate from ($f = 1$) and pressure to mutate toward ($f = 0$) a particular residue/base. The HKY model introduces another free parameter, γ, the transition to transversion ratio. For both DNA and protein models, we calculate the values of r, f, γ which maximize the likelihood of A given the tree.

For each of the match state models we use the emission frequency distribution of the corresponding match state as the steady-state frequency π. We use HMMER to calculate the profile HMM, and so we know that this distribution is a smoothed version of the raw column in the seed alignment. We note here the possibility of using techniques other than Dirichlet priors to smooth the column frequencies. One potential method is a tree-based smoothing technique (Qian and Goldstein, 2003; Mitchison, 1999) where the smoothed distribution is calculated as the posterior distribution at the root given the residues/bases observed at the leaves of the tree. We use the values of r, f calculated for the protein model, but note that it may be possible—by training on the column of the alignment which was used to build the profile HMM—to modify r, f to more accurately reflect evolutionary pressures at a given match state. We also note that in general it does not suffice to use the maximum likelihood values of r, f obtained from training on this column—this results in parameter overfitting. To rectify this, it will be necessary to introduce a prior distribution on r, f centered on the values calculated with the protein model. However, we have not pursued this option further.

An important advantage of the PSILC algorithms over dN/dS is the directionality of the calculations. For example, the score on an alignment of two transcripts $x_{1.}, x_{2.}$ is not symmetric: the PSILC score for $x_{1.}$ will not in general be equal to the PSILC score for $x_{2.}$. This is in contrast to the dN/dS calculation, which is symmetric. This directionality property of the PSILC calculation is due to the directionality of the underlying substitution models above:

$$\mathbf{P}_k(x_{1i}, x_{2i}, t) = \mathbf{P}_k(x_{2i}, x_{1i}, t) * \frac{\pi_{x_{1i}}}{\pi_{x_{2i}}} \neq \mathbf{P}_k(x_{2i}, x_{1i}, t) \qquad (9)$$

$$\text{unless } \pi_{x_{1i}} = \pi_{x_{2i}},$$

using the condition of detailed balance. To see that this is appropriate, consider the case in which the residue x_{1i} is more likely than x_{2i} at a particular match state but equally likely under the protein model. From Equations (1) and (9) the score from this column for $x_{2.}$ being a pseudogene is higher than for $x_{1.}$ under PSILC$_{\mathrm{prot/dom}}$. Thus we expect that PSILC will be better able to discriminate which of two related genes is a pseudogene, and which is the functional copy. In principal, dN/dS can accomplish this task, but only by calculating a dN/dS ratio with a third reference gene x_3, whereas it is an innate property of the PSILC score.

SYSTEM AND METHODS

Test data

The manual annotation of human chromosome 6 (Mungall *et al.*, 2003) (NCBI34 human genome build), which can be obtained from http://www.vega.sanger.ac.uk, was used as the principal test set for the method, we shall call this the Vega set. Vega annotates both functional genes and pseudogenes, and as such is an ideal test set. In general, Vega pseuodgenes are categorized on the basis of homology to known genes/proteins with a disrupted ORF due to frameshifts and/or inframe stop codons. Vega contains 1887 coding transcripts on chromosome 6 and 633 pseudogenes. Of these, we extracted 875 coding transcripts and 158 pseudogenes which could be aligned to at least one different ENSEMBL transcript using the protocol described below. Of these, we then extracted 598 (68%) coding transcripts and 97 (61%) pseudogenes which matched a Pfam domain. Pfam release 10.0 was used in this study.

Method

For each (pseudo)gene transcript in the test set a blast search against the ENSEMBL (Birney *et al.*, 2004) NCBI34 transcripts for human, rat and mouse was carried out. The query transcript and ENSEMBL transcripts with blast match *e*-value less than 10^{-7} and a cumulative match length greater than 80% of the query transcript were aligned. Transcripts with greater than 99% match on more than 80% of the original sequence were removed from the alignment, to avoid the inclusion of sequences from ENSEMBL which are effectively the same regions in Vega. The transcripts were aligned using CLUSTALW version 1.83. Columns in the alignment with stop codons or an incomplete codon (due to a frameshift in the pseudogene) were removed. A neighbor joining tree was calculated for the alignment with pairwise distances calculated as the maximum likelihood distances [using the PAL package (Drummond and Strimmer, 2001)] based on the DNA alignment and the HKY model of nucleotide evolution (Hasegawa *et al.*, 1985). Each Pfam family which was homologous to all of the transcripts in the alignment (using the Pfam annotation in ENSEMBL) was identified, and the profile HMM was aligned to the transcript alignment.

The tree, alignment and aligned Pfam domain form the inputs for the PSILC algorithm. For PSILC$_{\mathrm{nuc/dom}}$ and PSILC$_{\mathrm{prot/dom}}$, if multiple Pfam domains matched the alignment, or if a single Pfam domain matched multiple times then the scores under each of the models was added together to obtain a single score. This is justified by the fact that Pfam families do not overlap on these sequences, and the assumption of conditional independence of each of the domain matching regions in the query transcript given the rest of the alignment.

The dN/dS score was calculated on the full extent of the alignment. PAML codeml was used to calculate dN/dS, in a manner following the pseudogene survey in Torrents *et al.* (2003). If the brother node of the query node in the tree was also a leaf-node, then the sub-alignment of the query transcript and the transcript at the brother node was extracted, and PAML analysis was performed on this sub-alignment. Otherwise, the sequence at the brother node was reconstructed as the consensus sequence of all leaf-nodes below the brother node. Classifying genes on the basis of high dN/dS score is potentially sub-optimal, as genes undergoing positive selection are expected to have a high dN/dS score, and so will falsely be classified as pseudogenes. To avoid this, we also investigate classifying on the basis of abs[$\log(\mathrm{dN}/\mathrm{dS})$] as well as on the dN/dS score. This modified score will be close to 0 when dN/dS is close to 1 (which is the value expected for a pseudogene).

RESULTS

Figure 2 shows the receiver operating curve for PSILC and dN/dS on the Vega chromosome 6 test set. Table 1 shows the areas under the curve for each method. We see that PSILC$_{\mathrm{prot/dom}}$ performs better than all the other methods at most thresholds, and has the greatest area under the curve. PSILC$_{\mathrm{nuc/dom}}$ out-performs both dN/dS variants at false acceptance rates below 4% and at false acceptance rates 33–90%, and also has greater area under the curve than both dN/dS variants. Thresholding on dN/dS appears to be more successful in general than on abs[$\log(\mathrm{dN}/\mathrm{dS})$], which in effect implies that when using dN/dS for pseudogene classification it is best not to correct for putatively positively selected genes.

We note that our method only applies when a Pfam domain can be aligned to the transcript. In the Vega test set, this was possible for 68%/61% of coding transcripts/pseudogenes which could be aligned to a distinct ENSEMBL transcript.

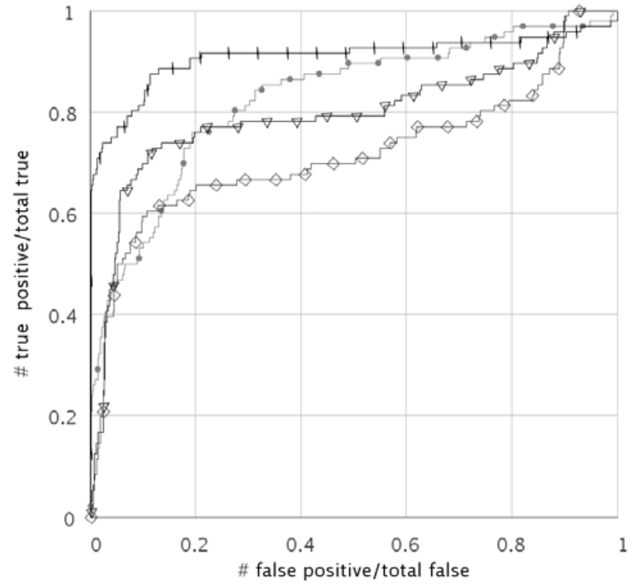

Fig. 2. Receiver operating curve for PSILC and dN/dS. Vertical dashes: $\text{PSILC}_{\text{prot/dom}}$; solid dots: $\text{PSILC}_{\text{nuc/dom}}$; inverted triangles: dN/dS; diamonds: abs[log(dN/dS)]. A larger area under the ROC represents a better discrimination between true and false pseudogenes. A classifier which picks pseudogenes at random would (on average) result in the line $x = y$.

Table 1. Area under the ROC for the different methods

PSILC-prot/dom	91%
PSILC-nuc/dom	83%
dN/dS	80%
abs[log(dN/dS)]	72%

A higher number represents a better overall classification.

This proportion is lower for pseudogenes as expected due further divergence away from the Pfam consensus.

It is remarkable that $\text{PSILC}_{\text{prot/dom}}$ out-performs all the other methods, as we are discarding all the information regarding synonymous to non-synonymous substitutions when performing this calculation. This demonstrates that divergence from a Pfam domain is a useful predictor of loss of functionality for a gene. We would expect that $\text{PSILC}_{\text{nuc/dom}}$, which combines information about divergence away from a Pfam domain with divergence away from protein coding sequence would out-perform $\text{PSILC}_{\text{prot/dom}}$. That this is not the case suggests that we are somehow penalizing DNA evolution relative to protein evolution in our protocol. It may be that the protein models are fitting the alignment data better (due to a higher degree of parameterization) and so despite choosing maximum likelihood parameters for the DNA model we are intrinsically penalizing DNA evolution.

Figure 3 shows the fraction of (pseudo)genes scoring above threshold versus threshold for both the $\text{PSILC}_{\text{prot/dom}}$ score and dN/dS ratio. The dN/dS graph is plotted on a log x-axis for clarity—the PSILC scores are effectively already log based scores. The dN/dS pseudogene distribution is centered on $\text{dN}/\text{dS} \approx 1$ as expected, and at $\text{dN}/\text{dS} \approx 0.1$ for functional genes. However, both distributions are spread over a large range of dN/dS values, which makes a clean separation on this score difficult. On the other hand, the functional genes have a much shaper distribution under the $\text{PSILC}_{\text{prot/dom}}$ score, with

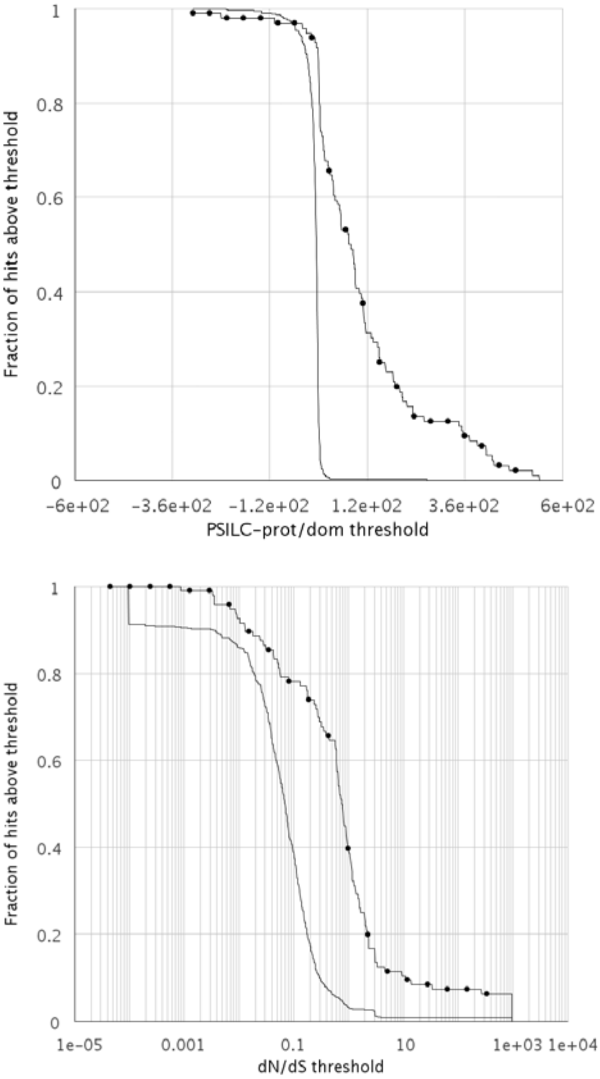

Fig. 3. Comparison of discrimination between pseudogenes and functional genes between the $\text{PSILC}_{\text{prot/dom}}$ method (top graph) and dN/dS (lower graph). In both graphs we plot the fraction of (pseudo)genes scoring above a particular threshold, with the pseudogenes represented by the line marked with dots, and functional genes represented by the line without dots. We see that the $\text{PSILC}_{\text{prot/dom}}$ method provides a cleaner separation threshold, at a threshold of zero (which reflects the log-odds nature of this scoring methodology—a score of greater than zero reflects more evidence in favor of pseudogene status than against).

most of the weight located at $\text{PSILC}_{\text{dom/prot}} \approx 0$, and the pseudogene distribution has most of its weight greater than 0, making a clean separation more effective.

DISCUSSION

We have demonstrated—with $\text{PSILC}_{\text{prot/dom}}$—the viability of a method which identifies pseudogenes without any knowledge of the rates of synonymous and non-synonymous substitution of the gene. Moreover, where a Pfam domain can be aligned with a gene, this method has been shown to be more accurate than dN/dS. Despite the limitation imposed by the requirement of detectable homology to a Pfam domain, this method is still of wide applicability (60–70% of sequences to which we can apply dN/dS).

There are several large-scale analyses for which the approach outlined in this paper would be useful. The first is a quality check on the gene annotation databases, such as ENSEMBL, to identify potential pseudogenes in these databases which are annotated as functional genes, and to identify genes annotated as pseudogenes which are more likely to be functional genes. The second is a scan of various genomes for pseudogenes, following Torrents *et al*. (2003) and Harrison *et al*. (2003). Using $\text{PSILC}_{\text{prot/dom}}$, it is also possible to perform an analysis of the functional DNA constraints on pseudogenes [as observed in Balakirev and Ayala (2003) for specific *Drosophila* genes] without the ascertainment bias of using the lack of functional DNA constraints to identify pseudogenes.

The approach outlined in this paper could be extended in several ways. It would be possible to infer loss of constraint along an entire clade of a tree, not just a final branch. Hence it is possible to test an entire clade for pseudogene status, and to identify the internal node of a tree at which the pseudogene arose. The method can also be used to score mutations (resulting from, e.g. SNPs) to predict the potential loss of functionality from a SNP. One potential problem with using $\text{PSILC}_{\text{prot/dom}}$ and/or $\text{PSILC}_{\text{nuc/dom}}$ for classifying pseudogenes is that genes under positive selection will be misclassified as pseudogenes. One way to resolve this issue is to develop a third score, $\text{PSILC}_{\text{nuc/prot}}$, which considers evolution away from protein coding in favor of neutral nucleotide evolution. The set of genes which have a strongly positive $\text{PSILC}_{\text{prot/dom}}$ score and a strongly negative $\text{PSILC}_{\text{nuc/prot}}$ score (hence are still evolving as proteins) may be an interesting candidate set for positive selection. If this were the case, then the method could also identify which domains in a multi-domain protein were undergoing selection.

ACKNOWLEDGEMENTS

We would like to thank Ewan Birney for suggesting this approach to identifying pseudogenes. We would like to thank Steven Kearnan, Jennifer Ashurst and Laurens Wilming for advice and help accessing the Vega gene annotation. We would also like to thank Val Curwen, Eduardo Eyers and Vivek Iyer for useful discussion and information relating to the ENSEMBL process for identifying pseudogenes, and for help accessing information in the ENSEMBL databases. The Wellcome Trust Sanger Institute is supported by the Wellcome Trust. LC holds a Leslie Wilson Scholarship from Magdalene College Cambridge and a Cambridge Australia Trust studentship.

REFERENCES

Andersson,J.O. and Andersson,S.G. (2001) Pseudogenes, junk, and the dynamics of Rickettsia genomes. *Mol. Biol. Evol.*, **18**, 829–839.

Balakirev,E.S. and Ayala,F.J. (2003) Pseudogenes: are they ‘junk’ or functional? *Annu. Rev. Genet.*, **37**, 123–151.

Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhoummer,E.L., *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.

Bierne,N. and Eyre-Walker,A. (2003) The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics*, **165**, 1587–1597.

Birney,E., Andrews,T.D., Bevan,P., Caccarno,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T., *et al*. (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.

Coin,L., Bateman,A. and Durbin,R. (2003) Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA*, **100**, 4516–4520.

Drummond,A. and Strimmer,K. (2001) PAL: An object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, **17**, 662–663.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Goldman,N. and Whelan,S. (2002) Novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.*, **19**, 1821–1831.

Goldman,N. and Yang,Z. (1994) Codon-based model of nucleotide substitution for protein-coding sequences. *Mol. Biol. Evol.*, **11**, 725–736.

Harrison,P.M. and Gerstein,M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.*, **318**, 1155–1174.

Harrison,P.M., Hegyi,H., Balasubramanian,S., Luscombe,N.M., Bertone,P., Echols,N., Johnson,T. and Gerstein,M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**, 272–280.

Harrison,P.M., Milburn,D., Zhang,Z., Bertens,P. and Gerstein,M. (2003) Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.*, **31**, 1033–1037.

Hasegawa,M., Kishino,H. and Yano,T. (1985) Dating of the human–ape splitting by a molecular clock of mitochondria. *J. Mol. Evol.*, **22**, 160–174.

Hirotsune,S., Yoshida,N., Chen,A., Garrett,L., Sugiyama,F., Takahashi,S., Yagami,K., Wynshaw-Boris,A. and Yoshiki,A.

(2003) An expressed pseudogene regulates the messenger stability of its homologous coding gene. *Nature*, **423**, 91–96.

Lio,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.

Mighell,A.J., Smith,N.R., Robinson,P.A. and Markham,A.F. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.

Mitchison,G.J. (1999) Probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.*, **49**, 11–22.

Mounsey,A., Bauer,P. and Hope,I.A. (2002) Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res.*, **12**, 770–775.

Mungall,A.J., Palmer,S.A., Sims,S.K., Edwards,C.A., Ashurst,J.L. Wilming,L., Jones,M.C., Horton,R., Hunt,S.E., Scott,C.E. *et al*. (2003) The sequence and analysis of human chromosome 6. *Nature*, **425**, 805–811.

Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Huhhard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Prince,V.E. and Pickett,F.B. (2002) Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.*, **3**, 827–837.

Qian,B. and Goldstein,R.A. (2003) Detecting distant homologs using phylogenetic tree-based hmms. *Proteins*, **52**, 446–453.

Torrents,D., Suyama,M., Zdobnov,E. and Bork,P. (2003) Genome-wide survey of human pseudogenes. *Genome Res.*, **13**, 2559–2567.

Whelan,S. and Goldman,N. (2001) General empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.