BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 24 2014, pages 3575–3582
doi:10.1093/bioinformatics/btu576

Sequence analysis

Advance Access publication August 28, 2014

Frameshift alignment: statistics and post-genomic applications

Sergey L. Sheetlin[1,†], Yonil Park[1,†], Martin C. Frith[2,*,†] and John L. Spouge[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA and [2]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The alignment of DNA sequences to proteins, allowing for frameshifts, is a classic method in sequence analysis. It can help identify pseudogenes (which accumulate mutations), analyze raw DNA and RNA sequence data (which may have frameshift sequencing errors), investigate ribosomal frameshifts, etc. Often, however, only *ad hoc* approximations or simulations are available to provide the statistical significance of a frameshift alignment score.

Results: We describe a method to estimate statistical significance of frameshift alignments, similar to classic BLAST statistics. (BLAST presently does not permit its alignments to include frameshifts.) We also illustrate the continuing usefulness of frameshift alignment with two 'post-genomic' applications: (i) when finding pseudogenes within the human genome, frameshift alignments show that most anciently conserved non-coding human elements are recent pseudogenes with conserved ancestral genes; and (ii) when analyzing metagenomic DNA reads from polluted soil, frameshift alignments show that most alignable metagenomic reads contain frameshifts, suggesting that metagenomic analysis needs to use frameshift alignment to derive accurate results.

Availability and implementation: The statistical calculation is available in FALP (http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/index/software.html), and giga-scale frameshift alignment is available in LAST (http://last.cbrc.jp/falp).

Contact: spouge@ncbi.nlm.nih.gov or martin@cbrc.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 12, 2014; revised on July 24, 2014; accepted on August 20, 2014

1 INTRODUCTION

Although the pairwise alignment of DNA and protein is a classical problem in bioinformatics (Guan and Uberbacher, 1996; Pearson *et al.*, 1997; Zhang *et al.*, 1997), it presents new challenges as next-generation sequencing (NGS) of DNA supplants Sanger sequencing. The per-base error of Sanger sequencing is as low as 0.001% (Shendure and Ji, 2008), but NGS can produce per-base errors of ~1–2% (Wang *et al.*, 2012). NGS can also make frequent insertion and deletion errors, e.g. 454 sequencing in HIV-1 produced occasional per-base insertion and deletion error rates of almost 50% (Shao *et al.*, 2013). More recent sequencing methods such as PacBio also have high indel rates (Carneiro *et al.*, 2012). The insertion and deletion errors cause frequent frameshifts when DNA is annotated by electronic translation and alignment to proteins. This article studies the statistics of 'frameshift alignment', a pairwise alignment of translated DNA and a protein, with particular attention to frameshift errors. Frameshift alignment has been used for protein domain classification (Zhang and Sun, 2011), determining exact gene start position (Baytaluk *et al.*, 2002), discovering distant protein homologies (Girdea *et al.*, 2010), predicting exon–intron structure (Mironov *et al.*, 2001) and so on.

Large post-genomic datasets require fast sequence comparison tools. There are several tools that can compare DNA sequences with proteins, including NCBI BLAST (Altschul *et al.*, 1990), USEARCH (Edgar, 2010), RapSearch (Zhao *et al.*, 2012), PAUDA (Huson and Xie, 2013) and GHOSTM (Suzuki *et al.*, 2012), but none of these consider frameshifts. The FASTA package has excellent support for frameshifts, but is not designed for large searches such as whole (meta)genomes versus all known proteins (Pearson *et al.*, 1997; Zhang *et al.*, 1997). LAST has supported frameshift alignment since 2009, and is designed for large searches (Kielbasa *et al.*, 2011). We should also mention HAXAT, which accurately aligns Roche 454 DNA sequences to proteins allowing for frameshifts, but is not designed for large-scale searches (Lysholm, 2012).

Some studies have focused on developing frameshift alignment algorithms; few have focused on the accurate estimation of E-values for frameshift alignment scores. In alignments without frameshifts, errors in approximate E-values can usually be controlled (Altschul *et al.*, 2001; Bundschuh, 2002; Park *et al.*, 2009); with frameshifts, errors are rarely controlled. On one hand, FASTX/Y calculates alignment *P*-values with z-scores (Pearson *et al.*, 1997). On the other hand, BLAST tools comparing translated DNA to protein do not consider frameshifts. Rather, they consider six alignment reading frames, calculating their E-values as though only one reading frame was explored (Gertz *et al.*, 2006; Gish and States, 1993). With NGS supplanting Sanger sequencing, the need for accurate frameshift alignment *P*-values has increased because users require an E-value accounting for frequent frameshifts to decide which sequences are true homologs.

The Ascending Ladder Program (ALP) was developed for real-time computation of the statistical parameters in BLAST (Park *et al.*, 2009, 2012; Sheetlin *et al.*, 2005). It is based on an importance sampling method using a global alignment of two simulated sequences generated by a Markov chain (Bundschuh, 2002; Park *et al.*, 2009). Heuristic modeling of the ascending alignment scores of the global alignment as a Markov additive

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors

process yielded analytical formulas for estimating the parameters of a modified Gumbel distribution.

This article has two aims: (i) it presents the Frameshift Ascending Ladder Program (FALP), which extends ALP by handling frameshift errors from NGS; (ii) it exemplifies the use of FALP by finding pseudogenes in the human genome and analyzing metagenomic DNA from polluted soil. FALP uses the same probabilistic framework as ALP, but the importance sampling for generating random sequence pairs requires seven Markov chain states in FALP, instead of three states in ALP.

The organization of this article is as follows. The Methods section describes the mathematics of frameshift alignment and our development and implementation of the FALP algorithm. It also describes heuristic frameshift alignment with LAST for genomic sequence data applications and provides the dataset we used for pseudogene and metagenome analyses. The Results section compares the approximate *P*-values from FALP to frameshift alignment *P*-values from simulation. It then shows how frameshift alignment helps to detect pseudogenes in the human genome and to analyze metagenomic DNA reads. Finally, the Discussion section reflects on our results. The Supplementary Data details the importance sampling used in FALP.

2 METHODS

2.1 Frameshift alignment and its notations

Let $\mathbf{A} = a_1 a_2 \ldots$ be a DNA sequence and $\mathbf{B} = B_1 B_2 \ldots$ be an amino acid sequence, where a_i is drawn from the nucleotide alphabet $\mathfrak{A}_A = \{a, c, g, t\}$, and B_j is from the amino acid alphabet $\mathfrak{A}_B = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Let $T(a_{i-2}a_{i-1}a_i) = A_i \in \mathfrak{A}_B^* = \mathfrak{A}_B \cup \{*\}$ be an amino acid translated from the codon $a_{i-2}a_{i-1}a_i$ or a stop codon denoted by $*$. Let $s : \mathfrak{A}_B^* \times \mathfrak{A}_B \mapsto \mathbb{R}$ denote a 'scoring matrix'. In database applications, $s(A, B)$ quantifies the similarity between A and B, e.g. the so-called 'PAM' (point accepted mutation) or 'BLOSUM' (block sum) scoring matrices can quantify evolutionary similarity between two amino acids (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992).

Let $w_g = \Delta_0 + \Delta_1 . g$ ($\Delta_0, \Delta_1 \geq 0$) denote the affine gap penalty, where Δ_0 is called the gap opening penalty, Δ_1 is called the gap extension penalty and g is the gap length in amino acids. Let γ be the 'frameshift penalty', imposed when sequencing errors necessitate a frameshift in the alignment of DNA and amino acid sequences. The scoring matrix $s(A, B)$, the gap penalty w_g and the frameshift penalty γ constitute the 'frameshift alignment scoring system'.

The frameshift alignment graph $\Gamma_{\mathbf{A},\mathbf{B}}$ of the sequence-pair $(\mathbf{A}, \mathbf{B})$ is a directed weighted lattice graph in two dimensions, as follows. The vertices v of $\Gamma_{\mathbf{A},\mathbf{B}}$ are non-negative integer points (i, j) (Throughout the article, i, j, k, m, n and g are integers.). Seven sets of directed edges e come out of each vertex $v = (i, j)$: one northward, three in slightly varying directions northeastward and three eastward. The first northeastward edge goes into $v = (i+3, j+1)$ with weight $s[e] = s(A_{i+3}, B_{j+1})$, corresponding to a match between a codon and an amino acid. The second northeastward edge goes into $v = (i+4, j+1)$ with weight $s[e] = s(A_{i+4}, B_{j+1}) - \gamma$, corresponding to a nucleotide insertion between two consecutive DNA codons. The third northeastward edge goes into $v = (i+2, j+1)$ with weight $s[e] = s(A_{i+2}, B_{j+1}) - \gamma$, corresponding to a nucleotide deletion, so that the last position of a codon overlaps with the first position of the next DNA codon. For each $g > 0$, one eastward edge goes into $v = (i+3g, j)$ and one northward edge goes into $v = (i, j+g)$, both assigned the same weight, $s[e] = -w_g < 0$. The remaining eastward edges go into $v = (i+3g+1, j)$ and $v = (i+3g-1, j)$ with weight $s[e] = -w_g - \gamma < 0$.

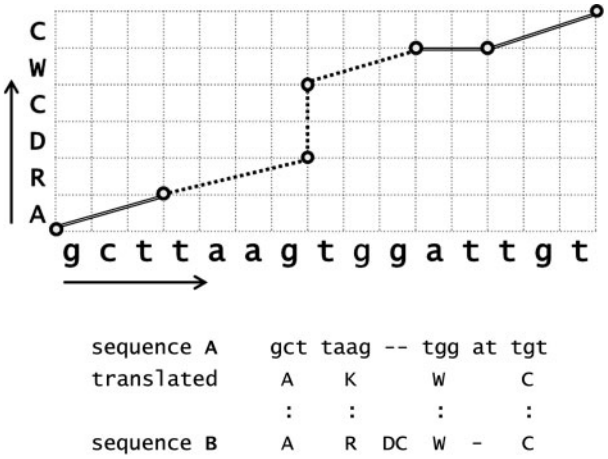

Fig. 1. Frameshift alignment graph for two sequences $\mathbf{A} = \texttt{gcttaagtggattgt}$ and $\mathbf{B} = \texttt{ARDCWC}$. Open circles indicate vertices v, and the origin $v_0 = (0, 0)$ is at the southwest corner of the Figure. The optimal frameshift alignment path consists of six edges, in order: 2 northeast, 1 north of length 2, 1 northeast, 1 east of length 1 and 1 northeast. It corresponds to the alignment word-pairs (gct, A), (taag, R), (Δ, D), (Δ, C), (tgg, W), (at, Δ) and (tgt, C). The frameshift alignment score is $\sum_{i=1}^{6} s[e_i]$, where $s[e_1] = s(\text{A}, \text{A})$, $s[e_2] = s(\text{K}, \text{R}) - \gamma$, $s[e_3] = -w_2$, $s[e_4] = s(\text{W}, \text{W})$, $s[e_5] = -w_1 - \gamma$ and $s[e_6] = s(\text{C}, \text{C})$. Note that frameshift occurs twice in the alignment by a nucleotide t insertion between gct and aag and a nucleotide deletion between tgg and tgt. Double line edges show that alignment is in reading frame 1 and dotted line edges show that alignment is in reading frame 2

A (directed) path $\pi = (v_0, e_1, v_1, e_2, \ldots, e_k, v_k)$ in $\Gamma_{\mathbf{A},\mathbf{B}}$ is a finite alternating sequence of vertices and edges that starts and ends with a vertex. For each $i = 1, 2, \ldots, k$, the directed edge e_i comes out of vertex v_{i-1} and goes into vertex v_i. We say that the path π starts at v_0 and ends at v_k and the alignment's score is the 'path weight' $S_\pi := \sum_{i=1}^{k} s[e_i]$. See Figure 1 for an example.

Let $\hat{M}_{m,n}$ denote the local maximum scores for a given DNA sequence $\mathbf{A} = a_1 a_2 \ldots a_m$ and amino acid sequence $\mathbf{B} = B_1 B_2 \ldots B_n$. In the appropriate limit, if the frameshift alignment scoring system is in the so-called 'logarithmic phase' (Arratia and Waterman, 1994; Dembo *et al.*, 1994) (i.e. if the optimal global alignment score of long random sequences has a negative score), the random score $\hat{M}_{m,n}$ follows an approximate Gumbel extreme value distribution with 'scale parameter' λ and 'pre-factor' K:

$$\mathbb{P}\left(\hat{M}_{m,n} \geq y\right) \approx 1 - \exp\left[-Kmn\exp\left(-\lambda y\right)\right]. \quad (1)$$

The search space size mn in Eq (1) is sharpened by edge effect (Altschul and Gish, 1996; Park et al., 2012).

2.2 Algorithm for frameshift alignment score

We assume that a nucleotide insertion or deletion error occurs only between two consecutive codons. Then, for gap penalty $w_g = \Delta_0 + \Delta_1 . g$ and frameshift penalty γ, the global frameshift alignment score $S_{i,j}$ is calculated with the recursion

$$S_{i,j} = \max \left\{ \begin{array}{l} \max\left\{S_{i-3,j-1}, D_{i-3,j-1}, I_{i-3,j-1}\right\}, \\ \max\left\{S_{i-2,j-1}, D_{i-2,j-1}, I_{i-2,j-1}, S_{i-4,j-1}, D_{i-4,j-1}, I_{i-4,j-1}\right\} - \gamma \end{array} \right\} + s(A_i, B_j), \quad (2)$$

$$I_{i,j} = \max\left\{S_{i,j-1} - \Delta_0, I_{i,j-1}\right\} - \Delta_1, \quad (3)$$

$$D_{i,j} = \max\left\{\max\left\{S_{i-3,j} - \Delta_0, D_{i-3,j}\right\}, \max\left\{S_{i-2,j} - \Delta_0, D_{i-2,j}, S_{i-4,j} - \Delta_0, D_{i-4,j}\right\} - \gamma\right\} - \Delta_1, \quad (4)$$

with boundary conditions $S_{0,0} = 0$, $D_{3g,0} = I_{0,g} = -\Delta_0 - \Delta_1.g$, $D_{3g-1,0} = D_{3g+1,0} = -\Delta_0 - \Delta_1.g - \gamma$, $S_{-1,0} = S_{-1,g} = S_{0,g} = S_{1,g} = S_{2,g} = S_{g,0} = -\infty$, $D_{-1,0} = D_{-1,g} = D_{0,0} = D_{0,g} = D_{1,0} = D_{1,g} = D_{2,g} = -\infty$, $I_{-1,0} = I_{-1,g} = I_{0,0} = I_{g,0} = -\infty$ for $g > 0$.

The three array names S, I and D are mnemonics for 'substitution', 'insertion' and 'deletion'. If 'Δ' denotes a gap character, the alignment word-pairs of the forms (aa, B), (aaa, B), $(aaaa, B)$, (Δ, B), (aa, Δ), (aaa, Δ) and $(aaaa, \Delta)$ correspond to the operations for editing sequence **A** into sequence **B** (Waterman *et al.*, 1976). Note that (aa, B) and (aa, Δ) represent a nucleotide deletion while $(aaaa, B)$ and $(aaaa, \Delta)$ represent a nucleotide insertion into the boundary of a codon.

The preceding calculation disallows adjacent insertions and deletions. If desired, they can be allowed by replacing the recursion for $I_{i,j}$ in Eq (3) with $I_{i,j} = \max\left\{S_{i,j-1} - \Delta_0, I_{i,j-1}, D_{i,j-1} - \Delta_0\right\} - \Delta_1$. As usual, the global alignment algorithm and its boundary conditions can be converted into a local alignment algorithm for comparing DNA and amino acid sequences.

2.3 Trial distribution for importance sampling

To estimate the Gumbel parameters, FALP uses importance sampling together with Eqs (2–4). Like all Monte Carlo methods, importance sampling estimates the integral of some function over some target distribution. It improves its estimate's accuracy by generating samples from a biased trial distribution that focuses its probability in a region where the function is large. It then corrects for sampling bias by multiplying the function values it samples by the ratio of target to trial probabilities before averaging. In the present context, our importance sampling generates frameshift alignments with high scores, so the Gumbel approximation is relatively accurate, to improve the accuracy of the corresponding estimated statistical parameters (Hammersley and Handscomb, 1964).

Let the letters in sequences **A** and **B** be independent and identically distributed with background (target) probabilities $\left\{p_a : a \in \mathfrak{A}_A\right\}$ and $\left\{q_B : B \in \mathfrak{A}_B\right\}$, respectively. In FALP, importance sampling generates high-scoring random sequence pairs with a trial distribution using a Hidden Markov model whose underlying Markov chain has seven states $\{S_1, S_2, S_3, D_1, D_2, D_3, I\}$, where $S_1 := \mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A \times \mathfrak{A}_B$, $S_2 := \mathfrak{A}_A\mathfrak{A}_A \times \mathfrak{A}_B$, $S_3 := \mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A \times \mathfrak{A}_B$, $D_1 := \mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A \times \{\Delta\}$, $D_2 := \mathfrak{A}_A\mathfrak{A}_A \times \{\Delta\}$, $D_3 := \mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A\mathfrak{A}_A \times \{\Delta\}$ and $I := \{\Delta\} \times \mathfrak{A}_B$. Let $t_{r,r'}$ represent a transition probability from state r to state r', where $r, r' \in \{S_1, S_2, S_3, D_1, D_2, D_3, I\}$. The transition probabilities $\left\{t_{r,r'}\right\}$ are determined using the scores $s(A, B)$, the gap penalty $w_g = \Delta_0 + \Delta_1.g$ and the frameshift penalty γ. The emission probabilities on D_1, D_2, D_3 and I states have the forms $p_ap_ap_a$, p_ap_a, $p_ap_ap_ap_a$ and q_B, respectively. The emission probabilities on S_1 and S_2 states have the forms $e_{aaa,B}$ and $e_{aa,B}$, respectively. The emission probabilities for the S_3 state have the form $p_ae_{aaa,B}$. The Supplementary Data has further details on the transition and emission probabilities. Under the assumption of independent letters in **A** and **B**, our methods then produce accurate statistical parameters applying to the output of the frameshift alignment algorithm in Section 2.2. As we generate sequence pairs by the hidden Markov model, we compute importance sampling weights as the ratio of the target probability over the corresponding trial probability. The importance sampling weights can be computed efficiently by a dynamic programming algorithm, similar to ALP (Park *et al.*, 2009), but with some technical complications, as described in the Supplementary Data.

2.4 Implementation

FALP can be run in three modes. 'Gumbel Mode' estimates the modified Gumbel parameters for a given frameshift alignment scoring system. The modified Gumbel parameters include the two parameters λ and K of the Gumbel distribution in Eq (1), together with parameters for a 'finite-size correction' to the Gumbel approximation. The finite-size correction accounts for edge effects caused by finite sequence lengths (Park *et al.*, 2012). '*P*-values Mode' calculates *P*-values and E-values using the modified Gumbel parameters. 'Align Mode' computes the optimal frameshift alignment score for a given pair of nucleotide and amino acid sequences. The default frameshift alignment scoring system in FALP is BLOSUM80 with $w_g = 11 + 2g$ and $\gamma = 15$. The default letter frequencies are uniform frequencies for the 4 nucleotides (all 0.25) and the Robinson and Robinson (1991) frequencies for the amino acids. Users can assign their own frameshift alignment scoring system and letter frequencies, and translation table.

FALP automatically allocates computation resources to compute all statistical parameters to the accuracy a user requires. The default accuracy for K is 0.5%. (Typically, K takes more time to estimate than λ.) FALP computes standard errors for its estimates of the modified Gumbel parameters by partitioning the random sequence-pairs from importance sampling into subsets. The user can specify the number of partition subsets, with default 10.

2.5 Fast heuristic frameshift alignment with LAST

For genomic-scale sequence data like genomes and metagenomes, we use LAST to find alignments quickly but heuristically. LAST follows the same three steps as BLAST: find 'seeds' (short initial matches); extend a gapless alignment from each seed; and if there is a sufficiently high-scoring gapless alignment, then extend a gapped alignment with frameshifts. LAST achieves adequate speed for genome-scale data by using adaptive seeds in the first step (Kielbasa *et al.*, 2011). In the third step, the frameshift recursion is calculated in a limited region of the dynamic programming matrix, defined by an x-drop algorithm (Altschul *et al.*, 1997). It uses this variant of the recursion, which aims to be as fast as possible (and assumes that $\gamma \geq \Delta_1$):

$$\begin{aligned} x &= \max\left\{S_{i-3,j-1}, S_{i-2,j-1} - \gamma, S_{i-4,j-1} - \gamma\right\} \\ y &= Y_{i,j-1} - \Delta_1 \\ z &= Z_{i-3,j} - \Delta_1 \\ b &= \max\left\{x, y, z\right\} \\ S_{i,j} &= b + s(A_i, B_j) \\ Y_{i,j} &= \max\left\{b - \Delta_0, y\right\} \\ Z_{i,j} &= \max\left\{b - \Delta_0, z\right\} \end{aligned}$$

This variant does consider adjacent insertions and deletions.

2.6 Pseudogene analysis

We identified human pseudogenes by comparing the genome (hg19) to a database of known proteins [UniRef90 version 2012_07, (Suzek *et al.*, 2007)]. The comparison was done with LAST v362 (Kielbasa *et al.*, 2011), after masking tandem repeats with tantan v13 (Frith, 2011). First, the proteins were masked and indexed:

```
tantan -p -r0.02 | lastdb -p -c
```

Then, the genome was masked and aligned:

```
tantan -c | lastal -pBLOSUM80 -F15 -e131 -m100 -f0
```

As a control, we also aligned the genome after reversing (but not complementing) it:

```
tantan | reverse | lastal -pBLOSUM80 -F15 -e93 -m100 -f0
-u3
```

The results were compared with the following annotation files from the UCSC database (Meyer et al., 2013): knownGene.txt, pseudoYale60.txt (Zhang et al., 2003), vegaPseudoGene.txt (Harrow et al., 2012). They were also compared with human-versus-lamprey alignments (hg19.petMar2.net.axt).

We also attempted a BLAST search, using NCBI BLAST+ 2.2.28. First, the proteins were masked and prepared:

```
tantan -p -r0.02 | makeblastdb -dbtype prot
```

Then, the genome was masked and aligned:

```
tantan -c |
blastx -matrix BLOSUM80 -outfmt 7 -lcase_masking
```

2.7 Metagenome analysis

We also analyzed a set of metagenomic DNA reads (SRR629686) by comparing it with UniRef90. The reads were first converted to fasta format, then masked and aligned as follows:

```
tantan | lastal -pBLOSUM80 -F15 -e129 -f0
```

3 RESULTS

3.1 FALP *P*-value evaluation

Table 1 gives FALP's estimates for scale parameter λ and prefactor K for the BLOSUM62 and $w_g = 11 + g$, the default scoring system in BLASTP for proteins. We used uniform nucleotide frequencies (all 0.25) and Robinson and Robinson (1991) frequencies for amino acids. As the frameshift penalty γ increases, estimates of λ converge to 0.330 ± 0.001. Figure 2 displays *P*-value accuracies for FALP, if the frameshift alignment scoring systems are BLOSUM62, $w_g = 11 + g$ and $\gamma = 9$, and BLOSUM80, $w_g = 11 + 2g$ and $\gamma = 15$. The 'gold standard' *P*-values were calculated using importance sampling in 2×10^8 simulation runs. Other scoring schemes also gave qualitatively similar results.

Table 1. Estimates for Gumbel parameters for BLOSUM62 and $w_g = 11 + g$

γ	λ	K
5	0.091 ± 0.004	0.00008 ± 0.00004
10	0.302 ± 0.002	0.120 ± 0.011
15	0.318 ± 0.003	0.132 ± 0.013
20	0.331 ± 0.002	0.186 ± 0.026
40	0.332 ± 0.002	0.099 ± 0.010

3.2 Finding pseudogenes in the human genome

Pseudogenes are gene fossils that no longer produce functional proteins, so they accumulate arbitrary mutations including frameshifts. One useful application of frameshift alignment is to identify pseudogenes by finding alignments (indicating homology) between a genome and all known proteins (from a protein database such as UniRef). The results will include exons of real genes as well as pseudogenes, but overlap with gene annotations identifies the real exons, which can then be removed. Gene annotations are often incomplete, so some real exons might remain, but for some applications this does not matter (see below).

Accordingly, we used frameshift alignment to find human pseudogenes. In fact, any frameshift aligner reasonably approximating LAST's algorithm above could have been used to produce the input scores for FALP. As usual, the alignment score threshold should be as low as possible, to find subtle pseudogenes, but not so low that it encourages 'too many' spurious alignments. FALP estimates that, if we compared a shuffled human genome to a shuffled UniRef90, with the default frameshift alignment scoring scheme, about 700 random alignments are expected with score $S \geq 100$, about 20 alignments with score $S \geq 110$, etc. (Table 2). We conservatively chose a score threshold of $S = 131$, with corresponding E-value $E \leq 0.01$, so almost no spurious alignments are expected.

Thus, FALP permits control of the number of spurious alignments in random sequences, but the implications of this control

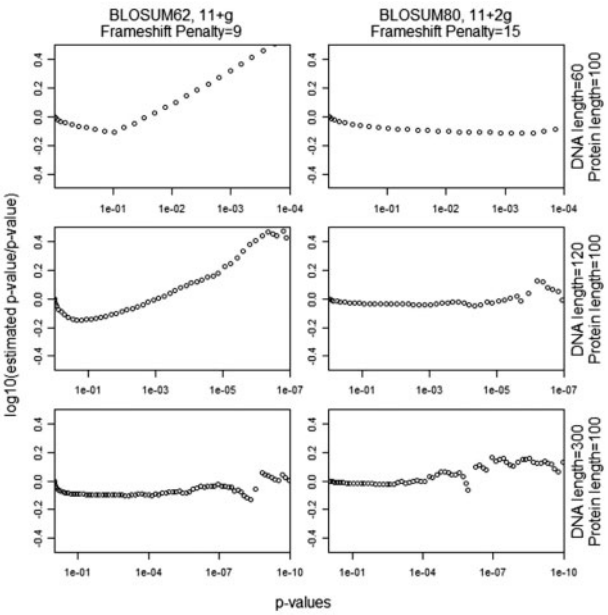

Fig. 2. Accuracy of FALP's *P*-value estimates. It plots $\log 10(\hat{p}/p)$ against p, where $\hat{p}$ is the FALP *P*-value and p is the simulation *P*-value. X-axis is in logarithmic scale. The first column plots are obtained from BLOSUM62 with $w_g = 11 + g$ and $\gamma = 9$. The second column plots are from BLOSUM80 with $w_g = 11 + 2g$ and $\gamma = 15$. The first row plots are obtained from a DNA sequence length 60 and protein sequence length 100; the second row DNA length 120 and protein length 100; the third row DNA length 300 and protein length 100. Perfect *P*-value estimation lies in $y = 0$

Table 2. Expected numbers of random alignments between the human genome and UniRef90

Minimum alignment score	Expected number of random alignments
90	26 538
100	714
110	19
120	0.52
130	0.014

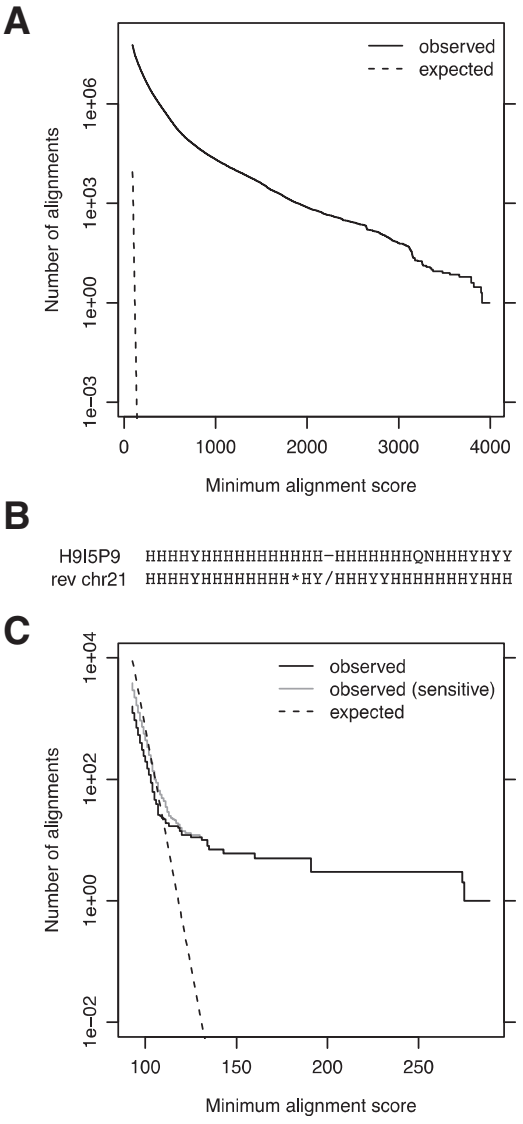

Fig. 3. LAST alignments between the reversed (but not complemented) human genome and known proteins in UniRef90. (**A**) Without repeat-masking. (**B**) An example of an alignment found without repeat-masking. (**C**) With repeat-masking (by tantan). The 'sensitive' line shows alignments found by running LAST in a more slow and sensitive mode

for spurious alignments in real DNA is unclear. To investigate, we aligned the reversed (but not complemented) human genome against UniRef90. DNA never evolves by reversal (due to its chemical 5′ to 3′ directionality), and so any alignments found are spurious (non-homologous). Figure 3A shows that FALP predicts far fewer spurious alignments than actually occur. Figure 3B suggests that matches between simple repeats account for most of the spurious alignments.

Figure 3C shows the results after masking simple repeats with tantan. FALP (dashed line) predicts the observed number of spurious alignments (black line) after repeat-masking with tantan, except for a dozen or so spurious alignments with unexpectedly large scores. All of the spurious alignments we investigated seem to arise from database artifacts, UniRef proteins that were computationally translated from reversed (but not complemented) RNA sequences. For example, UniRef proteins Q13540 and Q61506 are translations from human mRNA M14624 and mouse mRNA X00951, respectively, and the corresponding mRNAs match the reference genome when reversed (but not complemented).

Once the database artifacts in Figure 3C are discounted, the observed number of spurious alignments was slightly less than the expected number, simply because our heuristic alignment tool (LAST) misses some alignments. Indeed, when we reran LAST more slowly and sensitively (-m1000 instead of -m100), the observed number of spurious alignments increased (but did not exceed the expected number). In summary, after repeat-masking with tantan, FALP permits reliable control of the number of non-homologous alignments. [Note: this is not the case for repeat-masking programs other than tantan because they miss some cryptic simple repeats (Frith, 2011).]

Our pseudogene search (which took 13 CPU hours) identified 65 005 putative pseudogene parts, segments of the human genome aligning to known proteins but not overlapping with annotated protein-coding sequences. (We say 'parts' because one pseudogene may have several separate fragments.) As noted above, some of the putative pseudogene parts are probably real exons lacking annotations, but 36 376 of 65 005 had frameshifts, which real exons should lack.

We compared our putative pseudogene parts with two pseudogene catalogs: Yale pseudogenes (Zhang *et al.*, 2003) and Vega pseudogenes (Harrow *et al.*, 2012). (In these catalogs, one pseudogene may have multiple 'exons', i.e. genomic segments.) Of our 65 005 putative pseudogene parts, 43 554 (including 19 679 with frameshifts) have no overlap with any pseudogene in Yale and Vega, whereas our search missed only 4792 Yale and Vega exons. The detailed breakdown in Table 3 shows that our

Table 3. Comparison of our pseudogene parts with those in the Yale and Vega pseudogene catalogs

Pseudogenes	Exons/parts	Missed by Yale	Missed by Vega	Missed by LAST
Yale	23 321	0	9960	1768
Vega	20 588	7318	0	3276
LAST	65 005	46 739	50 222	0

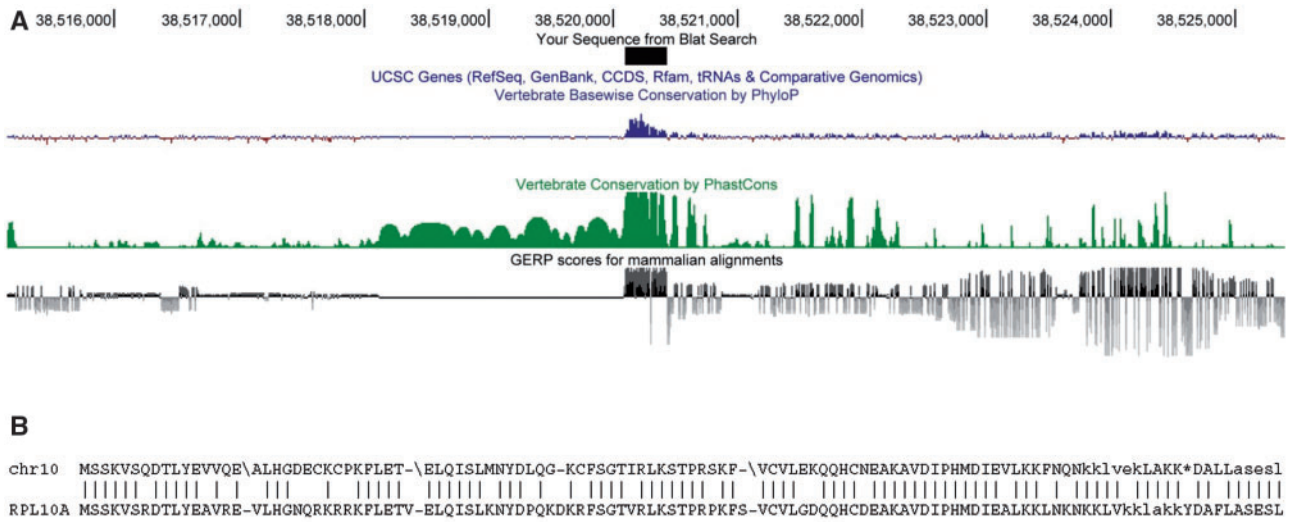

Fig. 4. Example of a conserved non-coding element that is a pseudogene. (**A**) Genomic context. The element itself is labeled 'your sequence from Blat search'. There are no annotated genes in this region. Three measures of conservation are shown: PhyloP, PhastCons and GERP. (**B**) Alignment of this DNA element (top, translated in three frames) to 60S ribosomal protein L10a

Table 4. Expected numbers of random alignments between a set of DNA reads (SRR629686) and a protein database (UniRef90)

Minimum alignment score	Expected number of random alignments
90	5996
100	195
110	6
120	0.21
130	0.0067

search missed a greater proportion of exons from Vega than from Yale, possibly because Vega includes transcribed pseudogenes, including exons not derived from protein-coding sequence.

To check whether our alignment methods have real benefit, we also attempted a BLAST search of the human genome against UniRef90. Using just chromosome 1, BLAST ran for 48 h with no output before it was killed by our queuing system. Thus, the mere ability to handle this scale of data is a non-trivial benefit.

3.3 Most anciently conserved non-coding elements are pseudogenes

Genomes contain conserved non-protein-coding elements, which often coincide with enhancers that appear to control fundamental aspects of embryonic development, including development of brain regions (Matsunami and Saitou, 2013; McEwen *et al.*, 2009). Therefore, studying these elements is a promising approach to understanding development. On the other hand, some non-coding elements are conserved for more prosaic reasons: recent pseudogenes can appear conserved simply because their parent genes were highly conserved.

Thus, to find the 'interesting' (regulatory) elements, we need to filter out the 'uninteresting' ones. To explore this issue, we examined alignments between the human and lamprey genomes, obtained from the UCSC database. Lamprey is a jawless vertebrate that is very distantly related to human, so these alignments reflect very ancient conserved elements. They include 25 383 putative non-coding elements, i.e. segments of the human genome that align to lamprey but have zero overlap with protein-coding sequence of known genes. Most of these non-coding elements (13 755: 54%) overlap pseudogene parts from our earlier search. This provides a straightforward (non-regulatory) explanation for these elements: presumably they became pseudogenes (via duplication or simply inactivation) much more recently than the common ancestor of human and lamprey, and their parent genes encoded conserved proteins.

One example is shown in Figure 4. This element appears strikingly conserved, and it lies in a gene desert (where conserved regulatory elements often lie). Without further information, we might suspect that it performs some fundamental, perhaps regulatory, function. Our alignments show, however, that it is actually a ribosomal protein pseudogene.

This is a case where it does not matter that our pseudogenes may include unannotated real exons. We wish to filter out conserved elements that have a straightforward explanation (that they either used to or still do encode proteins), to get the interesting unexplained ones.

3.4 Application to analyzing metagenomic DNA reads

By comparing metagenomic DNA reads to a protein database, we may gain clues as to what kinds of proteins and enzymes are encoded in the DNA (Darling *et al.*, 2014). As an example, we examined a set of DNA reads from industrially polluted soil in India (Shah *et al.*, 2013). This has 409 782 reads (133 529 997 bases) sequenced with the FLX titanium system. This sequencing system is often used in metagenomics because it produces fairly

```
     V   S   A   P   L   R   Q   A   L   P   E   G   S   S   L   V   F   !   I   H   A   T   E   P
     V   S   E   P   V   R   A   A   L   P   E   G   S   S   L   V   F   !   I   N   A   T   E   P
GGTT GTT TCG GAA CCT GTG CGC GCA GCC CTG CCC GAA GGC AGC TCC CTG GTT TTT C ATC AAT GCC ACG GAG CCT

 F   F   T   Y   I   K   L!S   A   M   A   G   L   L   L   S   L   P   V   I   F   W   Q   L   W
 F   F   T   Y   L   K   L!G   A   L   A   G   F   L   V   S   L   P   V   I   L   W   Q   I   W
TTT TTC ACC TAT CTC AAA CTGGG GCC TTG GCC GGC TTT CTC GTT TCC CTG CCC GTC ATC CTC TGG CAA ATC TGG G
```

Fig. 5. Example of a metagenomic DNA read aligned to a protein. The lower sequence is DNA read SRR629686.75083 (reverse strand). The upper sequence is part of a previously known protein: tatC, a sec-independent protein translocase, from *Pelobacter carbinolicus* (Q3A8D5). The middle sequence shows the translation of each codon, and frameshifts are indicated with '!'

long reads, which is especially beneficial for characterizing DNA from unknown or novel organisms. On the other hand, it has a significant rate of insertion and deletion errors (in particular, homopolymer length errors), which are likely to introduce frameshifts.

As usual we need to choose an alignment score threshold, so as to find as many significant alignments as possible, but not too many random alignments. Table 4 shows expected numbers of random alignments if we compare all the DNA reads to the protein database. (Note these are different from BLAST E-values, which are for comparing *one* query sequence to the database.) We conservatively used a minimum alignment score of 129, for an E-value of 0.01. (So the single-query E-value is ~0.01/409 782).

Our conservative alignment procedure found alignments for 254 768 (62%) of the reads. It took 142 CPU minutes [~15 real minutes, since we used multiple CPUs (Tange, 2011)]. Surprisingly, 145 181 reads (more than half of the aligned reads) have a top-scoring alignment that contains a frameshift. One example is shown in Figure 5. There are several possible explanations for seeing so many frameshifts:

- Frameshift errors in the DNA reads.
- Database errors (e.g. the database protein was deduced from a nucleotide sequence with frameshift errors).
- Non-erroneous DNA reads from pseudogenes.
- Evolution of proteins by frameshift.

This environment is polluted with chemical waste (from manufacture of dyes, paints, solvents, pharmaceuticals, etc.), likely including mutagens, which might conceivably cause some of the frameshifts. Although it is hard to distinguish sequence errors from real frameshifts, we can at least say that, since frameshifts are so frequent, it is important to take them into account to correctly understand relationships between metagenomic sequences and known proteins. For example, Figure 5 has two compensatory frameshifts, which means that non-frameshift alignment is likely to align the middle part wrongly, obscuring its evolutionary history.

4 DISCUSSION

We presented a new program FALP that assesses the statistical significance of a frameshift alignment score. In the algorithm section, we assumed that a nucleotide insertion or deletion occurs only between two consecutive codons. However, FALP can easily handle general error models because FALP is based on a Markov chain. Simply we define insertion or deletion error events as Markov states and add the corresponding Markov states to the current seven states in the Markov state-space.

We suggested a simple way of roughly annotating pseudogenes in a genome by frameshift alignment search of a protein database. These annotations are highly informative, even without tackling the hard question of distinguishing pseudogenes from real genes. In particular, these annotations overlap, and explain, most apparently conserved non-coding elements. This enables us to focus on the remaining unexplained conserved elements, which are likely to have important biological roles, such as regulation. On the other hand, 'ultraconserved elements' (Bejerano *et al.*, 2004; Lomonaco *et al.*, 2014) are not pseudogenes—see the Supplement.

We also analyzed one metagenomic DNA dataset by alignment to known proteins, and surprisingly found that most alignable reads exhibit frameshifts. This fact testifies to the importance of analyzing such data with frameshift alignments. Some frameshifts might not be sequencing errors, but could indicate interesting evolutionary events. Since frameshift alignment is rarely used, we suspect that there is a lot to discover with it.

Frameshift alignment may improve taxonomic identification of metagenomic sequences that lack close relatives in the database. This is because protein is typically more conserved than DNA, and so frameshift alignment should be able to find more distant relationships than the usual DNA-level alignment. It is also likely to improve gene prediction, where it is critical to get the correct reading frame. This study demonstrates that statistically rigorous large-scale frameshift alignment can be done easily, and we hope it will encourage renewed interest in this powerful technique.

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflict of interest: none declared.

REFERENCES

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Altschul,S.F. *et al.* (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.*, **29**, 351–361.

Arratia,R. and Waterman,M.S. (1994) A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, **4**, 200–225.

Baytaluk,M.V. *et al.* (2002) Exact mapping of prokaryotic gene starts. *Brief. Bioinformatics*, **3**, 181–194.

Bejerano,G. *et al.* (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Bundschuh,R. (2002) Rapid significance estimation in local sequence alignment with gaps. *J. Comput. Biol.*, **9**, 243–260.
Carneiro,M.O. *et al.* (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, **13**, 375.
Darling,A.E. *et al.* (2014) Phylosift: Phylogenetic analysis of genomes and metagenomes. *Peer J.*, **2**, e243.
Dayhoff,M.O. *et al.* (1978) A model of evolutionary change in proteins. In: *Atlas of protein sequence and structure*. Vol. Supp 3, pp. 345–352. National Biomedical Research Foundation, Silver Spring, MD.
Dembo,A. *et al.* (1994) Limit distributions of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, **22**, 2022–2039.
Edgar,R.C. (2010) Search and clustering orders of magnitude faster than blast. *Bioinformatics*, **26**, 2460–2461.
Frith,M.C. (2011) A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.*, **39**, e23.
Gertz,E.M. *et al.* (2006) Composition-based statistics and translated nucleotide searches: improving the tblastn module of blast. *BMC Biol.*, **4**, 41–41.
Girdea,M. *et al.* (2010) Back-translation for discovering distant protein homologies in the presence of frameshift mutations. *Algorithms Mol. Biol.*, **5**, 6.
Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.*, **3**, 266–272.
Guan,X.J. and Uberbacher,E.C. (1996) Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.*, **12**, 31–40.
Hammersley,J.M. and Handscomb,D.C. (1964) Monte Carlo methods. In: Bartlett,M.S. (ed.) *Monographs on Applied Probability & Statistics*. Methuen & Co.
Harrow,J. *et al.* (2012) Gencode: The reference human genome annotation for the encode project. *Genome Res.*, **22**, 1760–1774.
Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Huson,D.H. and Xie,C. (2013) A poor man's blastx—high-throughput metagenomic protein database search using pauda. *Bioinformatics*, **30**, 38–39.
Kielbasa,S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.*, **21**, 487–493.
Lomonaco,V. *et al.* (2014) UCbase 2.0: ultraconserved sequences database (2014 update). *Database*, **2014**, pii: bau062.
Lysholm,F. (2012) Highly improved homopolymer aware nucleotide-protein alignments with 454 data. *BMC Bioinformatics*, **13**, 230.
Matsunami,M. and Saitou,N. (2013) Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. *Genome Biol. Evol.*, **5**, 140–150.
Mcewen,G.K. *et al.* (2009) Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet.*, **5**, e1000762.
Meyer,L.R. *et al.* (2013) The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
Mironov,A.A. *et al.* (2001) Pro-Frame: similarity-based gene recognition in eukaryotic DNA sequences with errors. *Bioinformatics*, **17**, 13–15.
Park,Y. *et al.* (2012) New finite-size correction for local alignment score distributions. *BMC Res. Notes*, **5**, 286–286.
Park,Y. *et al.* (2009) Estimating the gumbel scale parameter for local alignment of random sequences by importance sampling with stopping times. *Ann. Stat.*, **37**, 3697–3714.
Pearson,W.R. *et al.* (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
Robinson,A.B. and Robinson,L.R. (1991) Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl Acad. Sci. USA*, **88**, 8880–8884.
Shah,V. *et al.* (2013) Taxonomic profiling and metagenome analysis of a microbial community from a habitat contaminated with industrial discharges. *Microb. Ecol.*, **66**, 533–550.
Shao,W. *et al.* (2013) Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in hiv-1 DNA. *Retrovirology*, **10**, 18.
Sheetlin,S. *et al.* (2005) The gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. *Nucleic Acids Res.*, **33**, 4987–4994.
Shendure,J. and Ji,H.L. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
Suzek,B.E. *et al.* (2007) Uniref: Comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, **23**, 1282–1288.
Suzuki,S. *et al.* (2012) Ghostm: a gpu-accelerated homology search tool for metagenomics. *Plos One*, **7**, e36060.
Tange,O. (2011) *GNU Parallel: The Command-Line Power Tool.;login: The USENIX Magazine*. pp. 42–47.
Wang,X.V. *et al.* (2012) Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, **13**, 185.
Waterman,M.S. *et al.* (1976) Some biological sequence metrics. *Adv. Math.*, **20**, 367–387.
Zhang,Y. and Sun,Y. (2011) Hmm-frame: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics*, **12**, 198.
Zhang,Z. *et al.* (1997) Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.*, **4**, 339–349.
Zhang,Z.L. *et al.* (2003) Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.*, **13**, 2541–2558.
Zhao,Y.A. *et al.* (2012) Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, **28**, 125–126.