BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2432–2439
doi:10.1093/bioinformatics/btu329

Sequence analysis

Advance Access publication May 13, 2014

SIBIS: a Bayesian model for inconsistent protein sequence estimation

Walyd Khenoussi, Renaud Vanhoutrève, Olivier Poch and Julie D. Thompson*

Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de médecine translationnelle, Strasbourg, F-67085, France

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The prediction of protein coding genes is a major challenge that depends on the quality of genome sequencing, the accuracy of the model used to elucidate the exonic structure of the genes and the complexity of the gene splicing process leading to different protein variants. As a consequence, today's protein databases contain a huge amount of inconsistency, due to both natural variants and sequence prediction errors.

Results: We have developed a new method, called SIBIS, to detect such inconsistencies based on the evolutionary information in multiple sequence alignments. A Bayesian framework, combined with Dirichlet mixture models, is used to estimate the probability of observing specific amino acids and to detect inconsistent or erroneous sequence segments. We evaluated the performance of SIBIS on a reference set of protein sequences with experimentally validated errors and showed that the sensitivity is significantly higher than previous methods, with only a small loss of specificity. We also assessed a large set of human sequences from the UniProt database and found evidence of inconsistency in 48% of the previously uncharacterized sequences. We conclude that the integration of quality control methods like SIBIS in automatic analysis pipelines will be critical for the robust inference of structural, functional and phylogenetic information from these sequences.

Availability and implementation: Source code, implemented in C on a linux system, and the datasets of protein sequences are freely available for download at http://www.lbgi.fr/~julie/SIBIS.

Contact: thompson@unistra.fr

Received on February 7, 2014; revised on April 4, 2014; accepted on May 5, 2014

1 INTRODUCTION

Next-generation sequencing (NGS) technologies are revolutionizing genomics, but the annotation of the genome assembly produced by these sequencers remains a major challenge (Guigo *et al.*, 2006; Yandell and Ence, 2012). The first step in many automatic genome annotation pipelines involves the prediction of protein-coding genes, a process that is intrinsically complicated, time-consuming and error-prone. First, DNA sequencing errors can lead to errors in gene sequence prediction, in particular those produced by NGS technologies or low-coverage assemblies (Hoff, 2009; Hubisz *et al.*, 2011; Trimble *et al.*, 2012). Second, the DNA errors are further confounded by inaccuracies in the methods used to delineate the protein-coding genes. Coding regions are mostly predicted by automatic methods, but the relationship between genes, transcripts and proteins is complex, and automated genome annotation is not completely accurate. The detection of protein-coding genes in prokaryotic genomes is generally considered relatively simple compared with eukaryotic genomes; however, there still remain a number of problems, including the detection of small genes (Warren *et al.*, 2010) or the localization of the start site (Gallien *et al.*, 2009; Venter *et al.*, 2011). In eukaryotic genomes, recent analyses have shown that the complete exon/intron structure is correctly predicted for only ~50–60% of genes (Brent, 2008; Guigo *et al.*, 2006; Harrow *et al.*, 2009). Even the best gene predictors and genome annotation pipelines rarely exceed accuracies of 80% at the exon level, meaning that most gene annotations contain at least one mis-annotated exon. The situation is further complicated by widespread alternative splicing events, which affect >92–94% of multi-exon human genes (Hallegger *et al.*, 2010).

Inconsistencies or inaccuracies in the prediction of protein-coding genes can often lead to errors in subsequent structural, functional or phylogenetic analyses. For example, it has been shown that low-coverage genomes generate not only a massive number of false gene losses but also striking artifacts in gene duplication inference (Milinkovitch *et al.*, 2010). It has also been demonstrated that contamination of databases with incomplete, abnormal or mispredicted sequences introduces a bias in the definition of orthologs (Dalquen *et al.*, 2013), the analysis of protein domain architectures between orthologs (Nagy *et al.*, 2011; Prosdocimi *et al.*, 2012) or the estimation of positive Darwinian selection (Schneider *et al.*, 2009). The accumulation of erroneous information in genomic and protein databases will continue to grow, as features are frequently transferred from annotated to unknown sequences (Gilks *et al.*, 2005), which only amplifies the level of errors in the databases. Such studies clearly highlight the urgent need for error detection and quality control strategies to reduce the impact of inconsistencies and to efficiently extract knowledge from the new genome data.

A number of computational methods have been developed to identify inconsistencies or errors in genome and protein sequences. At the genome level, current efforts in sequencing error mitigation mainly rely on filtering strategies, including filtering for sequencing read depth, base call quality, short-read alignment quality, variant call quality, known variants, strand bias, allelic imbalance and sequence context. All of these post-processing techniques help to reduce uncertainty in the final

*To whom correspondence should be addressed.

genotyping variant call (Robasky *et al.*, 2014). At the protein level, one approach toward solving this problem is to incorporate protein, EST and RNA-seq evidence that may support or contradict the exon/intron structure of the annotated gene. Several metrics have been developed, for example, the annotation edit distance (AED) (Eilbeck *et al.*, 2009), which measures how congruent each annotation is with its overlapping evidence. The MisPred method (Nagy and Patthy, 2013; Nagy *et al.*, 2008) is based on the hypothesis that conflicts in structural or functional annotations might imply a badly predicted sequence. Unfortunately, as stated by the authors, although MisPred identifies many suspicious sequences, it detects only a fraction of the truly erroneous sequences. We have previously developed an alternative approach that incorporates evolutionary information and does not require any structural or functional knowledge (Prosdocimi *et al.*, 2012; Thompson *et al.*, 2011). This approach was used to evaluate the effect of sequence errors on multiple sequence alignment accuracy (Thompson *et al.*, 2011), as well as in evolutionary and functional enrichment analyses (Prosdocimi *et al.*, 2012). In this method, the evolutionary conservation of a set of related sequences was measured using Gribskov profiles (Gribskov *et al.*, 1987), representing observed amino acid frequencies in alignment columns. Nevertheless, there are a number of shortcomings in estimating scores for amino acids based only on their observed frequencies, especially when the number of observed sequences is small. First, unseen amino acids are assigned scores based on a substitution matrix (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992) that is generally chosen arbitrarily. Second, an arbitrary threshold must be defined to distinguish related from unrelated or erroneous sequences.

Here, we present a new method, called SIBIS (Bayesian Inconsistency in Sequences), for the detection of sequence inconsistencies or errors, which replaces the Gribskov profiles used previously with a Bayesian model (Altschul *et al.*, 2010) to represent the evolutionary information in multiple sequence alignments and to estimate the relatedness of specific sequence segments. The Bayesian approach provides a strong theoretical foundation for modeling the amino acid frequencies found at a specific alignment position, and for calculating scores for including new sequences (Altschul *et al.*, 2010; Sjolander *et al.*, 1996; Ye *et al.*, 2011). SIBIS thus combines a prior distribution of amino acid probabilities, with observed amino acid frequencies at homologous positions within the related proteins. In theory, the prior distribution may be of any form, but here we use a powerful and elegant formalism, which is to assume that the amino acid frequencies follow a Dirichlet distribution or a mixture of a finite number of Dirichlet distributions (Sjolander *et al.*, 1996). We then exploit an efficient method for calculating posterior distributions, which was developed in the context of 'Bayesian Integral Log-odds' (BILD) substitution scores (Altschul *et al.*, 2010), designed to extend the log-odds formalism that is widely used to define residue substitution scores for pairwise alignments, to multiple alignments.

Similar Bayesian approaches have been used previously to improve the accuracy and sensitivity of various sequence analysis tools, including sequence database searches (Sjolander *et al.*, 1996) and in Gibbs sampling optimization procedures for local multiple alignment (Altschul *et al.*, 2010). Here, we use the Bayesian model to estimate the relatedness of individual sequences within a conserved alignment region. To do this, we compare the probabilities of observing specific sequences, under the assumptions of relatedness and unrelatedness. Unrelated sequence segments are then flagged as sequence inconsistencies or potential errors. The accuracy of our prediction is evaluated and compared with two existing methods, based on a set of eukaryotic sequences with known sequence errors that were validated experimentally (Zhang *et al.*, 2012). The results indicate that the Bayesian approach is statistically most powerful and has the highest accuracy for detecting inconsistent sequence segments, corresponding to ambiguous sequence variants or errors.

2 METHODS

2.1 Protein sequence test sets

To test the performance of the sequence error detection methods, we used a reference set of protein sequences with known errors that were validated experimentally (Zhang *et al.*, 2012). The protein sequences resulted from an automatic genome annotation of the draft genome sequence and assembly of the rhesus macaque (*Macaca mulatta*) (Gibbs *et al.*, 2007). The genome was sequenced with about a 5.2-fold coverage, and protein-coding genes were identified by comparison with human gene sequences. In a subsequent study (Zhang *et al.*, 2012), a number of rhesus macaque genes, including the first 100 genes of rhesus chromosome 20, were compared with the orthologous human genes to identify likely sequence errors. A number of the predicted errors were then validated by targeted re-sequencing. We excluded 10 gene sequences from the reference set that were not available in the sequence databases. Thus, the final reference set consisted of 90 protein sequences, of which 37 sequences had identified errors and 53 sequences were assumed to be correct. The sequence errors included examples of badly predicted N/C-terminal positions (43%), insertions/deletions (11%) and suspicious sequence segments (46%).

In addition, we constructed two independent test sets of protein sequences from the Uniprot database (Uniprot_Consortium, 2014). The first test set was constructed by randomly selecting 90 sequences from human chromosome 7, for which the protein names did not contain the words 'putative' or 'uncharacterized', and evidence for the protein's existence was available at the protein or transcript level. The sequences in this set were thus considered to be 'reliable'. The second test set consisted of 90 'unreliable' sequences selected using the following criteria: protein existence is 'predicted' (from the Uniprot sequence annotation), sequence status is 'complete' (i.e. the sequence is not known to be a fragment in Uniprot) and protein name contains 'putative uncharacterized'.

The complete list of protein sequences used in our experiments is available at http://www.lbgi.fr/~julie/SIBIS.

2.2 Construction of multiple alignments

For each protein sequence in the reference and test sequence sets, we searched for homologous sequences in the Uniprot database using BlastP (Altschul *et al.*, 1997). A multiple alignment of the top hits in the BlastP search (expect value <10, maximum 250 sequences) was constructed with the DbClustal program (Thompson *et al.*, 2000). DbClustal combines the advantages of both local and global alignment algorithms into a single system, and is thus able to provide accurate global alignments of the full-length protein sequences. For comparison purposes, we also constructed multiple alignments of the top hits found by BlastP searches in the Uniref90 and Uniref50 databases.

We identified potentially conserved regions in the multiple alignment using a strategy similar to that incorporated in RASCAL (Thompson *et al.*, 2003). Briefly, the sequences in the complete alignment were first

divided into more related subfamilies based on an automatically computed dissimilarity threshold implemented in the Secator program (Wicker *et al.*, 2001), and conserved 'core blocks' were identified for each subfamily. The core blocks, representing the sequence segments that are conserved in the majority of the sequences within the subfamily, were determined using the mean distance (MD) column scores implemented in the NorMD objective function (Thompson *et al.*, 2001). A sliding window analysis of the MD scores was performed and a threshold was defined above which columns were considered to be conserved (Thompson *et al.*, 2003).

These multiple alignments and the defined core blocks are used as input for the detection of inconsistent sequence segments, using either the Bayesian approach described below or an existing profile-based method (Prosdocimi *et al.*, 2012; Thompson *et al.*, 2011).

2.3 Prediction of inconsistent sequence segments

Given a conserved core block within a multiple alignment, we wanted to estimate the relatedness of a given sequence segment to the core block to predict unrelated or erroneous sequences. To do this, we used the Bayesian approach described by Altschul *et al.* (2010) for the calculation of BILD scores for each column in the core block. These scores are designed to extend log-odds scores for amino acids, such as those used in pairwise substitution matrices [Dayhoff PAM (Dayhoff *et al.*, 1978), Blosum (Dayhoff *et al.*, 1978) etc.], to multiple alignment columns. Here, we describe how we use this approach to calculate posterior probability distributions for amino acids in each column of the conserved core blocks, given a prior probability distribution and the observed frequencies of the amino acids in the alignment columns. The posterior probabilities are then used to identify sequence segments that are unrelated to the other sequences within the core block.

2.4 Dirichlet prior probability distributions

The prior probability distributions of the amino acids can be defined by Dirichlet distributions or Dirichlet mixtures. Given an amino acid alphabet of length L, a Dirichlet distribution has parameters α_1 to α_L, where $\alpha_j > 0$.

Because different regions of a protein sequence can have different evolutionary pressures, we use a Dirichlet mixture to define different prior probabilities for different collections of amino acids. A Dirichlet mixture is a set of standard Dirichlet distributions with parameters α_{i1} to α_{iL}, for the ith component Dirichlet distribution. Each component distribution has a coefficient m_i, where the m_i sum to 1.

To our knowledge, the only Dirichlet mixture prior parameters for protein sequence alignments have been derived by the team who first proposed such mixtures (Sjolander *et al.*, 1996), and these have been made available at compbio.soe.ucsc.edu/dirichlets/index.html. In the experiments described in the Section 3, we used a 20 component Dirichlet mixture (recode3.20comp) to define the prior probability distribution Θ_0 associated with a column of related amino acids. This mixture was derived from analyses of large numbers of alignments of related protein sequences, and has a relative entropy of 0.61, roughly equivalent to that of the PAM-175 matrix. In our experiments, this mixture achieved the highest sensitivity compared with the other distributions available (data not shown). General pointers for the choice of a suitable Dirichlet mixture prior have been discussed previously in Altschul *et al.* (2010) and at compbio.soe.ucsc.edu/dirichlets.

2.5 Bayes' theorem to derive posterior distributions

Posterior distributions can be obtained by modifying the prior probability distribution after observation of the amino acids in a given alignment column. To do this, we use the method of Altschul *et al.* (2010), which is reproduced here for the sake of completeness. For an alignment of M sequences, denote a column of residues: x_1 to x_M. Then we may apply Bayes' theorem to transform the prior distribution Θ_0 to a posterior Θ_1, after the observation of x_1. The posterior distribution Θ_1 is also a Dirichlet distribution α', where $\alpha'_x = \alpha_x + 1$, but with all other parameters unchanged. More generally, each subsequent observation x_k can be seen to transform the prior Θ_{k-1} into a posterior distribution Θ_k. A simplified example of the derivation of posterior distributions for an alphabet of three characters (D,I,L) is shown in Table 1, for an alignment column of two residues (I,L) and in Table 2, for an alignment column of (D,D).

Given the observation of residue x, and a Dirichlet mixture prior of C components, the parameters m'_i and α'_{ij} of the posterior distributions of amino acid probabilities may be calculated efficiently as follows:

For i = 1 to C:

(i) Multiply the mixture coefficient m_i by the Bayesian factors:

$$\tilde{m}_i = m_i \frac{\alpha_{i,x}}{\alpha_i^*}$$

Table 1. Simplified example of Bayesian inference of posterior distributions after observation of an alignment column containing residues (I,L)

Variable	Prior distribution: Θ_0		Distribution after observation (I): Θ_1		Distribution after observation (I,L): Θ_2	
	C 1	C 2	C 1	C 2	C 1	C 2
Coeff: m_i	0.60	0.40	0.15	0.85	0.11	0.89
Sum(α)	1.80	3.70	2.80	4.70	3.80	5.70
D	1.50	0.10	1.50	0.10	1.50	0.10
I	0.10	1.70	1.10	2.70	1.10	2.70
L	0.20	1.90	0.20	1.90	1.20	2.90
Prob(I|Θ_2)	0.22		0.55		0.45	
Prob(L|Θ_2)	0.27		0.35		0.49	
Prob(D|Θ_2)	0.51		0.10		0.06	

Note: Distributions are two-component Dirichlet mixtures (C = component, coeff = coefficient, prob = probability of observing a new residue).

Table 2. Example of Bayesian inference of posterior distributions after observation of an alignment column containing residues (D.D)

	Prior distribution: Θ_0		Distribution after observation (D): Θ_1		Distribution after observation (D.D): Θ_2	
	C1	C 2	C 1	C 2	C 1	C 2
Coeff: m_i	0.60	0.40	0.98	0.02	0.99	0.01
Sum(α)	1.80	3.70	2.80	4.70	3.80	5.70
D	1.50	0.10	2.50	1.10	3.50	2.10
I	0.10	1.70	0.10	1.70	0.10	1.70
L	0.20	1.90	0.20	1.90	0.20	1.90
Prob(I|Θ_2)	0.22		0.04		0.03	
Prob(L|Θ_2)	0.27		0.08		0.05	
Prob(D|Θ_2)	0.51		0.88		0.92	

Note: Distributions are two-component Dirichlet mixtures (comp = component, coeff = coefficient, prob = probability of observing a new residue).

$$m'_i = \frac{\tilde{m}_i}{\sum_{i=1}^{C} \tilde{m}_i}$$

(ii) Normalize the sum of the mixture coefficients to 1:
(iii) Add 1 to each $\alpha_{i,x}$: For j from 1 to L, $\alpha'_{ij} = \alpha_{ij} + 1$ if j = x, and $\alpha'_{ij} = \alpha_{ij}$ otherwise.

For small numbers of sequences, the probability distributions resemble the Dirichlet mixture prior, while for large numbers of sequences, the probability distributions converge toward the observed frequencies.

2.6 Estimation of related sequence segments

The posterior distributions Θ_M after observing the alignment column x_1 to x_M can be used to calculate the probability of observing a new residue x_{M+1}, under the assumption of relatedness:

$$Prob(x|\Theta) = \sum_{i=1}^{c} m_i \frac{\alpha_{ix}}{\sum_{j=1}^{L} \alpha_{ij}}$$

As an example, the probability of observing a residue I, given an alignment column of two residues (I,L) as shown in Table 1, is Prob(I) = $0.11 \times 1.1/3.8 + 0.89 \times 2.7/5.7 = 0.45$.

As the sequence segments are only evaluated in the conserved core block regions, gaps are not expected to occur often. Therefore, we assign the probability of observing a gap an arbitrarily low value of 10^{-2}.

The score for a segment of length N aligning to the model (under the simplifying assumption that each position in the protein is generated independently) is equal to the product of the probabilities of aligning each residue to the corresponding column in the model:

$$SegmentScore = \prod_{k=1}^{N} Prob(x_{M+1}|\Theta_M)$$

Thus, in the example shown in Table 1 and 2 of an alignment consisting of two columns (I,L) and (D,D), the score for aligning a related sequence 'ID' = $0.45 \times 0.92 = 0.41$, and the score for aligning an unrelated sequence 'DL' = $0.06 \times 0.05 = 0.003$.

To estimate the probability of observing a sequence segment under the assumption of unrelatedness, we calculate the score of a random sequence equal to the length of the core block with background amino acid frequencies equal to 1/L. Although this clearly does not reflect natural amino acid abundances, the use of more realistic frequencies would require a more time-consuming simulation of a large set of random sequences. Finally, sequence segments with a score less than that obtained by the random sequence are flagged as inconsistent sequences. It should be noted that the use of equal background frequencies results in a higher probability for the random sequence and will lead to some 'inconsistent' segments being rejected.

2.7 Prediction of N/C-terminal sequence errors

The Bayesian approach described above can be used to estimate the relatedness of sequence segments to a conserved region of the alignment. To detect badly predicted insertions and deletions, alternative methods are used:

- Badly predicted start or stop sites are identified by considering the positions of the N/C-terminal residues for each sequence in the subfamily alignment. For each sequence, the position of the terminal residue in the alignment is noted. A window, W, of 'normal' values is then determined as follows: Q1–10 < W < Q3 + 10, where Q1 and Q3 are the lower and upper quartiles, respectively, of the distribution of terminal positions. Sequences with terminal positions outside this window are annotated as potential deletion/extension errors.
- Inserted sequence segments are detected when a segment of at least 10 unaligned residues belonging to a particular sequence are flanked by alignment blocks including all the sequences in the subfamily. More formally, a potential inserted segment is detected if two subfamily alignment columns (i.j) exist such that [($n_i = N_i$) AND ($n_j = N_j$) AND ($N_k = 1$ for $i < k < j$) AND ($j-i > = 10$)], where N_i is the total number of sequences in the subfamily (excluding fragments at column i), n_i is the number of residues in column i.
- Similarly, missing sequence segments are detected using the following rule: a potential missing exon is detected if two subfamily alignment columns (i.j) exist such that [($n_i = N_i$) AND ($n_j = N_j$) AND ($N_k = N-1$ for $i < k < j$) AND ($j-i > = 10$)], where N_i is the total number of sequences in the subfamily (excluding fragments at column i), n_i is the number of residues in column i.

The heuristics described in this section are applied in both the Bayesian-based method and the profile-based method (Prosdocimi *et al.*, 2012; Thompson *et al.*, 2011) used for comparison purposes.

2.8 Output of the algorithm

Sequence segments estimated to be unrelated or erroneous are stored as features in an XML output file (Thompson *et al.*, 2006) that provides an appropriate format for use in automatic sequence analysis pipelines, thus facilitating integration in high-throughput projects. Files are also generated for input to the JalView program (Waterhouse *et al.*, 2009) for editing and viewing sequence alignment annotations.

3 RESULTS AND DISCUSSION

3.1 Algorithm overview

To detect badly predicted protein sequences, we developed a new method that combines a general framework for the analysis of evolutionary conservation information from multiple sequence alignments (Thompson *et al.*, 2011) with theoretically robust Bayesian-based scores for the estimation of the relatedness of individual sequences (Altschul *et al.*, 2010). The method, called SIBIS, is outlined in Figure 1.

A multiple alignment of the protein sequences to be evaluated is needed as input. The sequences in the alignment are then clustered automatically into more closely related subfamilies and conserved 'core blocks' are defined. Within these core blocks, the aligned sequence segments are scored based on the probabilities of observing the letters in each alignment column under the assumption that the sequences are related, and under the assumption that they are unrelated (see Section 2 for details). The sequence segments estimated to be unrelated, or erroneous, are output in an XML format file, allowing automatic parsing of the results for use in automatic sequence analysis pipelines.

3.2 Evaluation and comparison with existing methods

We applied our method to a set of protein sequences predicted from the rhesus macaque (*M.mulatta*) genome. A previous study had identified errors in the first 100 genes of rhesus chromosome 20 and a number of the errors were then validated by targeted resequencing (Zhang *et al.*, 2012). We extracted a reference set of 90 protein sequences from this study, of which 37 sequences had erroneous segments and 53 sequences were assumed to be

correct. By comparing the results of our method with the reference set of badly predicted sequences, we estimated the accuracy of our approach in terms of sensitivity and specificity. We also compared the performance of our Bayesian-based method with two previously published algorithms. First, MisPred (Nagy *et al*., 2008; Nagy and Patthy, 2013) is based on the rationale that if the features of the predicted protein conflict with existing knowledge (e.g. violation of protein domain integrity or co-occurrence of extracellular and nuclear domains), then the gene sequence is likely to be mispredicted. Second, we used a method that we developed previously (Thompson *et al*., 2011), which is similar to SIBIS in that it incorporates information from multiple sequence alignments. However, the sequence conservation estimation is based on the construction of Gribskov profiles (Gribskov *et al*., 1987).

As shown in Figure 2, our method achieves a sensitivity of 81%, which is significantly higher than both MisPred (27%) and the profile-based method (62%) (McNemar's test, 1-tail $P = 0.00766$). The low sensitivity of Mispred may be owing to the fact that it requires external information, for example, about structural domains or specific functions, which is not available for all proteins. Nevertheless, some of the errors that were not detected by SIBIS were correctly identified by Mispred, and the two approaches could be considered to be complementary. The higher sensitivity of SIBIS compared with the profile-based method can be attributed directly to the more robust Bayesian estimation of sequence relatedness, as both of these methods incorporate evolutionary information from the same multiple alignments.

It should be noted that, by default, we used sequences from the Uniprot database to construct the multiple alignment input to SIBIS. However, the sensitivity of the SIBIS method (like the profile-based method) will depend on the similarity of the input sequences. For example, when the sequences in the alignment all share <90% identity (extracted from the Uniref90 database), the sensitivity of SIBIS is reduced to 75%, and when the sequences in the alignment all share <50% identity (extracted from Uniref50), the sensitivity is only 53%.

Figure 3 shows an example of an inconsistent sequence segment identified by SIBIS, which was not detected by the profile-based method. The macaque sequence XP_002802392.1 is annotated in the NCBI RefSeq database as a predicted WD repeat-containing protein 90-like sequence. The inconsistent segment, covering amino acids 805–834, does not match the human ortholog WDR90_HUMAN or any of the other closely related homologs in the alignment.

The slight loss of specificity (92%) of the SIBIS method, compared with the profile-based method (96%), is not statistically significant (McNemar's test, 1-tail $P = 0.5$). Furthermore, the true specificity of the methods is difficult to estimate accurately using this test set, as some erroneous sequence segments may not have been annotated in the original study. Errors were originally identified in the macaque sequences by comparing them manually to the orthologous human sequences. Only those sequences estimated to contain errors were then validated experimentally. Therefore, it is possible that some sequences that were assumed to be correct may in fact contain errors. To illustrate this point, the sequence XP_001084527.2 (Fig. 4) was identified as correct in Zhang *et al*. (2012), although the sequence has an N-terminal extension compared with the most closely related sequences in the alignment. XP_001084527.2 is annotated as 39S ribosomal protein L28, mitochondrial isoform 2, but there is no experimental evidence for the existence of this isoform. The Human orthologous sequence, XP_005255097.1, has also been predicted using an automatic gene annotation method.

These tests highlighted a number of other problems, including the prediction of the various isoforms coded by a gene and the

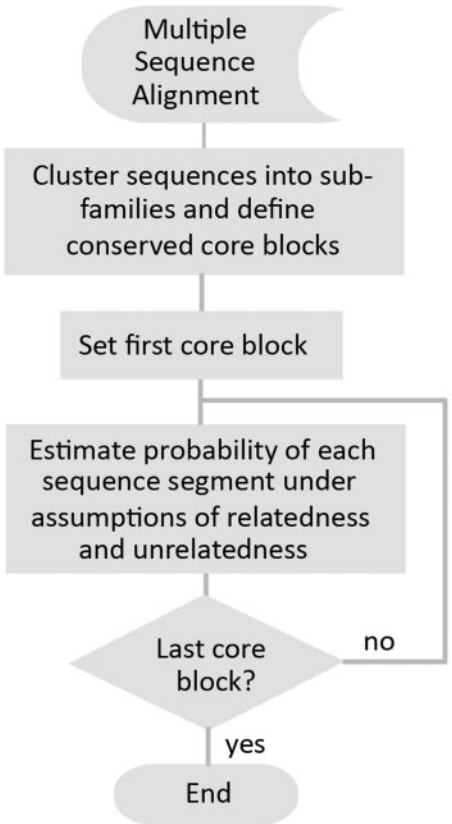

Fig. 1. Schematic diagram of the steps involved in the SIBIS protein sequence error detection method

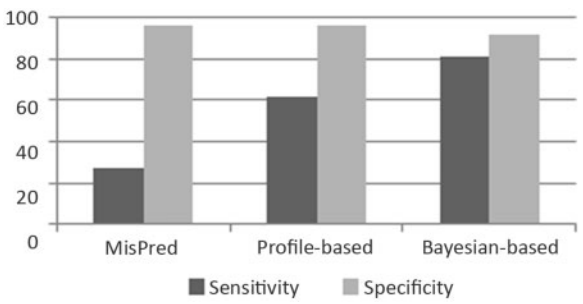

Fig. 2. Accuracy of three different methods for the detection of errors in protein sequences. Sensitivity and specificity were calculated based on a reference set of known errors in sequences predicted from the rhesus macaque genome (sensitivity = true positives/true positives + false negatives; specificity = true negatives/true negatives + false positives)

propagation of errors or inconsistencies to closely related organisms. The example alignment shown in Figure 5 illustrates these two issues. In this example, sequences detected by a BlastP search with the macaque protein XP_001087099.1 were aligned, including the sequence HAGHL_HUMAN (Hydroxyacylglutathione hydrolase-like protein), which is highly similar to the macaque sequence in the N-terminal region, but has a divergent C-terminus (Fig. 5A). It is important to note here that Swissprot selects a principal isoform for each gene, termed the reference sequence (or isoform 1). The reference sequence in the HAGHL_HUMAN Swissprot entry corresponds to an isoform that has not been identified in other organisms. There is no experimental evidence at the protein level for this isoform and a BlastP search using HAGHL_HUMAN as a query found no other sequences similar to the C-terminal segment. In fact, the predicted macaque sequence (XP_001087099.1), as well as the HAGHL reference sequences (isoform 1) in the mouse and chicken (HAGHL_MOUSE, HAGHL_CHICK), correspond to isoform 2 of Swissprot HAGHL_HUMAN (Fig. 5B).

The definition of different transcripts or gene products for different organisms can lead to misleading or false conclusions in subsequent analyses, for example, when identifying conserved residues or building phylogenetic trees, as shown in Figure 5C. Here, the macaque sequence is clustered with the mouse sequence (XP_001087099.1 shares 81% residue identity with HAGHL_MOUSE and only 73% with HAGHL_HUMAN), and based on this phylogenetic tree, it would be possible to infer an innovation in the human–chimpanzee lineage. The transcript problem exists not only in Swissprot but also in other sequence databases, such as Ensembl, where 11 isoforms are predicted for HAGHL_HUMAN and 6 for HAGHL_MOUSE. The isoform problem is a major issue because alternative splicing is common in eukaryotes, affecting 85% of protein-coding genes in humans for example (Rodriguez *et al.*, 2013), where it has been suggested to be a means of increasing protein complexity from a finite number of genes.

The example in Figure 5 also highlights the problem of error propagation, as the human isoform 1 has been used to model the H2RD11_PANTR chimpanzee sequence, for which only one

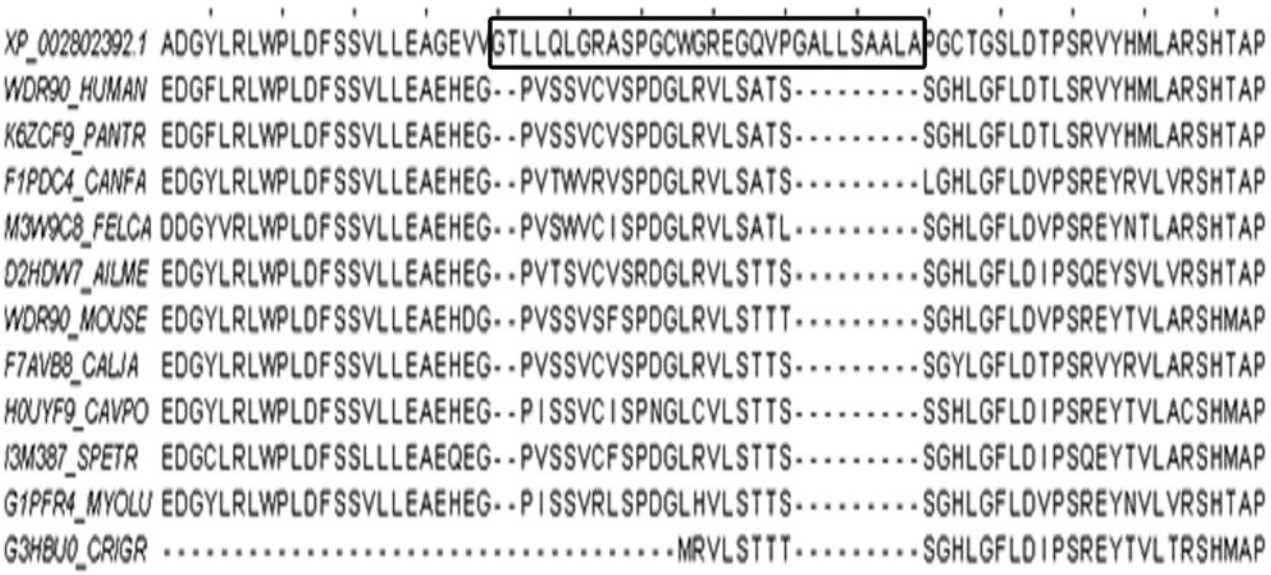

Fig. 3. Part of a multiple alignment of homologs of the rhesus macaque sequence XP_002802392.1. The black box indicates a predicted error in the N-terminal region of this sequence

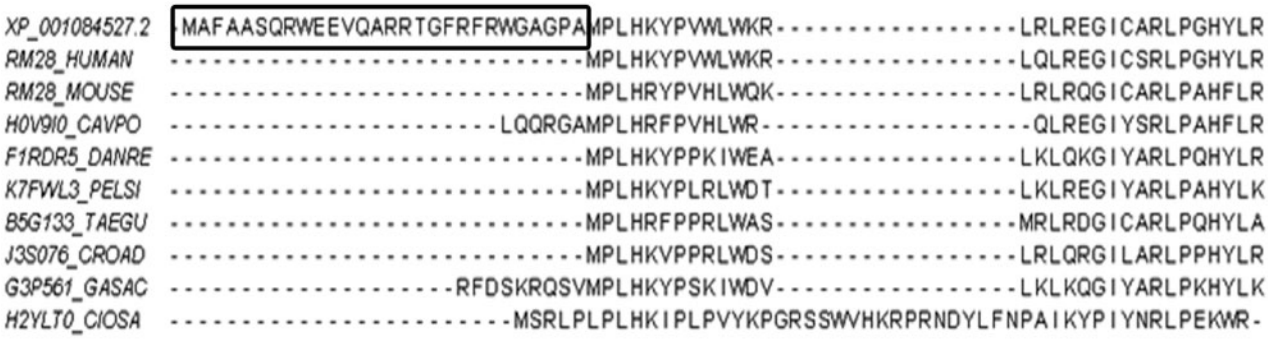

Fig. 4. Part of a multiple alignment of homologs of the rhesus macaque sequence XP_001084527.2. The black box indicates a predicted error in the N-terminal region of this sequence

isoform has been predicted corresponding to isoform 1 of HAGHL_HUMAN. When an error occurs in the identification of the exon/intron structure of a gene, the error clearly has consequences for the protein sequences of the studied organism. But, the consequences can be more wide-reaching because existing sequences in the public databases are often used to guide the gene-finding process in the annotation of new genomes.

3.3 Estimation of human protein error rates

We used the SIBIS method to evaluate the consistency of two test sets, each containing 90 human protein sequences from the Uniprot database. The 90 sequences in the first set were considered to be reliable (supported by evidence at the transcript or protein level), while the 90 sequences in the second set were potentially unreliable (protein existence is 'predicted' and protein name contains 'putative uncharacterized'). As for the previous dataset described in Section 3.2, a multiple alignment of homologs was constructed for each sequence in the test sets, and inconsistent sequence segments were predicted using SIBIS.

Inconsistencies were found in only one of the 90 reliable sequences, namely DLX6_HUMAN (Uniprot:P56179). In fact, the principal isoform, or reference sequence, in the Swissprot database corresponds to an isoform of length 175, which is expressed mainly in embryos. However, most of the other sequences found in the multiple alignment correspond to isoform 3 of DLX6_HUMAN, with length 293, with the exception of the cat and mouse proteins. This isoform is annotated in the CCDS database (Farrell *et al.*, 2014) as an alternative splicing pattern that is more supported by the available transcript and homology data. Although there are no full-length human transcripts spanning the first exon of this update, it is supported by tiled EST data and by mouse mRNA EF535989.1.

In contrast, 44 (48%) of the 90 unreliable sequences were found to contain at least one inconsistency/error. This level of error confirms previous estimates of the performance of gene-finding programs, which reached only ~40–50% accuracy at the transcript level (Brent, 2008; Guigo *et al.*, 2006; Harrow *et al.*, 2009).

4 CONCLUSION

In spite of tremendous advances in computational gene finding, comprehensive and robust genome annotation remains a challenge because of the exon/intron structure of eukaryotic genes and the complex relationship between genes, transcripts and proteins. As a consequence, computationally predicted genes and proteins should be confirmed by independent evidence and/or manual verification.

Experimental validation of protein sequences, for example, using shotgun proteomics, is one solution, but this is clearly infeasible for all proteomes. An alternative is to use bioinformatics approaches to identify erroneous, ambiguous or inconsistent protein sequences, using additional information such as EST sequences, protein domain families or evolutionary information. Here, we used multiple alignments to identify non-conserved or unrelated segments in sets of related sequences. To assess the evolutionary conservation between sequences in the alignment, we used a Bayesian approach to combine Dirichlet mixture distributions with the observed frequencies of the amino acids in the alignment columns. The Bayesian model provides a robust, theoretically sound representation of sequence relatedness, and, in our experiments, we have shown that it significantly improves the sensitivity of error detection, without loss of specificity. A major advantage of our method is that it does not assume that structural or functional information is available, so it can also be applied to sequences from less well-studied or uncharacterized organisms.

The interpretation of the inconsistencies detected using bioinformatics approaches remains an issue. There is a risk that a genetic event (recombination, alternative splicing,

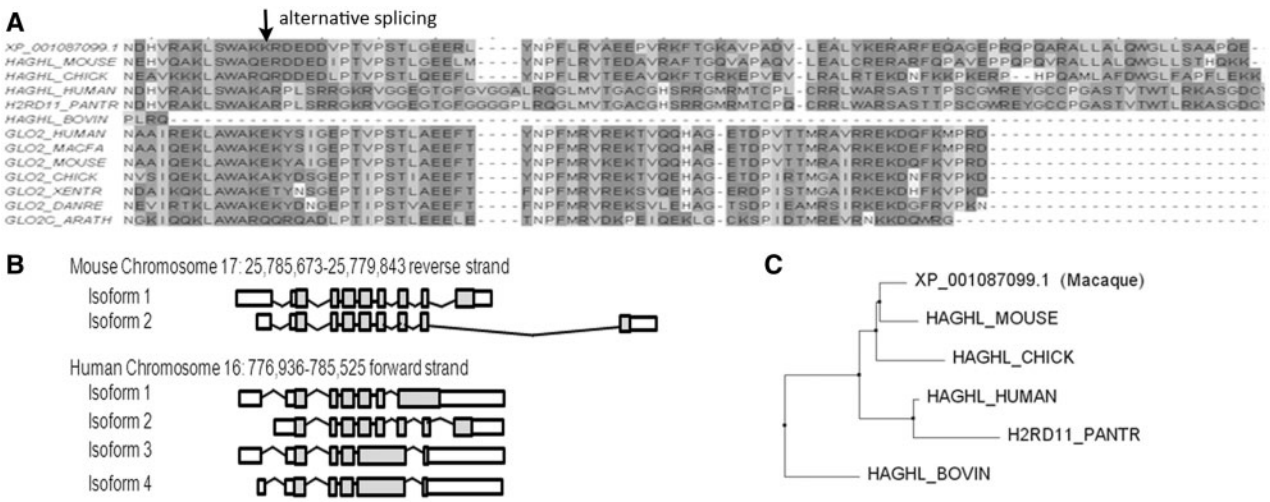

Fig. 5. (**A**) Part of a multiple alignment of homologs of the rhesus macaque sequence XP_001087099.1 with sequences from the Swissprot database. The arrow indicates a predicted alternative splice site in the C-terminal region of the human sequence HAGHL_HUMAN. (**B**) Exon structure corresponding to the isoforms of orthologous genes HAGHL_MOUSE and HAGHL_HUMAN annotated in the Swissprot database. (**C**) Phylogenetic tree of HAGHL family sequences, constructed using Mr. Bayes (Ronquist *et al.*, 2012)

pseudogenization etc.), which affects a single sequence in the multiple alignment, could be interpreted as an error. Nevertheless, the majority of the inconsistencies detected by SIBIS are likely to result from genome sequencing errors or inaccuracies in the gene annotation pipeline.

Our work confirms previous studies (Brent, 2008; Guigo *et al*., 2006; Harrow *et al*., 2009), which estimated that approximately half of the sequences in the public databases contain errors, and it is clear that simply eliminating these sequences from subsequent analyses is not a viable solution. Our method has the advantage that the errors are delimited, so that the reliable sequence segments can be used in subsequent studies. Error detection methods, such as the one described here, that are capable of accurately distinguishing between reliable and unreliable sequence segments will be crucial for automatic sequence analysis pipelines and should lead to more robust structural, functional and phylogenetic analyses.

ACKNOWLEDGEMENTS

The authors would like to thank Odile Lecompte and Luc Moulinier for helpful discussions and the ICube common services for their support.

Funding: This work was supported by the Agence Nationale de la Recherche (BIPBIP: ANR-10-BINF-03-02), the Région Alsace and Institute funds from the CNRS (Centre Nationale de Recherche Scientifique), the Université de Strasbourg and the Faculté de Médecine de Strasbourg.

Conflict of Interest: none declared.

REFERENCES

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*. (2010) The construction and use of log-odds substitution scores for multiple sequence alignment. *PLoS Comput. Biol*., **6**, e1000852.

Brent,M.R. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet*., **9**, 62–73.

Dalquen,D.A. *et al*. (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PLoS One*, **8**, e56925.

Dayhoff,M.O. *et al*. (1978) A model of evolutionary change in proteins. In: Foundation,N.B.R. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington DC, pp. 345–352.

Eilbeck,K. *et al*. (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, **10**, 67.

Farrell,C.M. *et al*. (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res*., **42**, D865–D872.

Gallien,S. *et al*. (2009) Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res*., **19**, 128–135.

Gibbs,R.A. *et al*. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.

Gilks,W.R. *et al*. (2005) Percolation of annotation errors through hierarchically structured protein sequence databases. *Math. Biosci*., **193**, 223–234.

Gribskov,M. *et al*. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Guigo,R. *et al*. (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol*., **7** (**Suppl. 1**), S2.1–S2.31.

Hallegger,M. *et al*. (2010) Alternative splicing: global insights. *Febs. J*., **277**, 856–866.

Harrow,J. *et al*. (2009) Identifying protein-coding genes in genomic sequences. *Genome Biol*., **10**, 201.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hoff,K.J. (2009) The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics*, **10**, 520.

Hubisz,M.J. *et al*. (2011) Error and error mitigation in low-coverage genome assemblies. *PLoS One*, **6**, e17034.

Milinkovitch,M.C. *et al*. (2010) 2x genomes–depth does matter. *Genome Biol*., **11**, R16.

Nagy,A. *et al*. (2008) Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics*, **9**, 353.

Nagy,A. and Patthy,L. (2013) MisPred: a resource for identification of erroneous protein sequences in public databases. *Database (Oxford)*, **2013**, bat053.

Nagy,A. *et al*. (2011) Reassessing domain architecture evolution of metazoan proteins: major impact of gene prediction errors. *Genes*, **2**, 449–501.

Prosdocimi,F. *et al*. (2012) Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics*, **13**, 5.

Robasky,K. *et al*. (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet*., **15**, 56–62.

Rodriguez,J.M. *et al*. (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res*., **41**, D110–117.

Ronquist,F. *et al*. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol*., **61**, 539–542.

Schneider,A. *et al*. (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol*., **1**, 114–118.

Sjolander,K. *et al*. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci*., **12**, 327–345.

Thompson,J.D. *et al*. (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*, **6**, e18093.

Thompson,J.D. *et al*. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*, **7**, 318.

Thompson,J.D. *et al*. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol*., **314**, 937–951.

Thompson,J.D. *et al*. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res*., **28**, 2919–2926.

Thompson,J.D. *et al*. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.

Trimble,W. *et al*. (2012) Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics*, **13**, 183.

Uniprot_Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*., **42**, D191–D198.

Venter,E. *et al*. (2011) Proteogenomic analysis of bacteria and Archaea: a 46 organism case study. *PLoS One*, **6**, e27587.

Warren,A.S. *et al*. (2010) Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, **11**, 131.

Waterhouse,A.M. *et al*. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wicker,N. *et al*. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol*., **18**, 1435–1441.

Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet*., **13**, 329–342.

Ye,X. *et al*. (2011) On the inference of dirichlet mixture priors for protein sequence comparison. *J. Comput. Biol*., **18**, 941–954.

Zhang,X. *et al*. (2012) Limitations of the rhesus macaque draft genome assembly and annotation. *BMC Genomics*, **13**, 206.