Published online 7 July 2007 Nucleic Acids Research, 2007, Vol. 35, No. 14 4737–4742
doi:10.1093/nar/gkm281

Intron length distributions and gene prediction

Scott William Roy* and David Penny

Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Received February 7, 2007; Revised March 21, 2007; Accepted April 11, 2007

ABSTRACT

Accurate gene prediction in eukaryotes is a difficult and subtle problem. Here we point out a useful feature of expected distributions of spliceosomal intron lengths. Since introns are removed from transcripts prior to translation, intron lengths are not expected to respect coding frame, thus the number of genomic introns that are a multiple of three bases ('3*n* introns') should be similar to the number that are a multiple of three plus one bases (or plus two bases). Skewed predicted intron length distributions thus suggest systematic errors in intron prediction. For instance, a genome-wide excess of 3*n* introns suggests that many internal exonic sequences have been incorrectly called introns, whereas a deficit of 3*n* introns suggests that many 3*n* introns that lack stop codons have been mistaken for exonic sequence. A survey of genomic annotations for 29 diverse eukaryotic species showed that skew in intron length distributions is a common problem. We discuss several examples of skews in genome-wide intron length distributions that indicate systematic problems with gene prediction. We suggest that evaluation of length distributions of predicted introns is a fast and simple method for detecting a variety of possible systematic biases in gene prediction or even problems with genome assemblies, and discuss ways in which these insights could be incorporated into genome annotation protocols.

INTRODUCTION

Ever since the dawn of the genomic age, accurate prediction of protein coding (and other) genes has been a central problem of biology (1,2). New annotations are continually released, even of the best-annotated and most carefully studied genomes. The statistical task of distinguishing true coding genes from non-coding sequences requires evading an array of pitfalls, among them pseudogenes, coding elements of repetitive sequences, short ORFs that are biologically but not statistically significant, and transcription of non-coding sequences. A wide variety of statistical and bioinformatic approaches to gene prediction have been developed, including comparisons with sequences from transcripts and from other species, statistical analysis of ORF length and Bayesian comparison of gene models (3–7).

One of the major obstacles to accurate gene prediction in eukaryotes is the presence of spliceosomal introns. Splicesomal introns are genomic sequences that are removed from RNA transcripts prior to translation by a very large RNA–protein complex called the spliceosome. Spliceosomal introns show large variations in intron number per gene, typically exhibit no length or sequence conservation either within or between species, and afford opportunities for alternative splicing, further complicating accurate deduction of a species' protein arsenal.

Here we point out a simple aspect of the expected distribution of spliceosomal intron lengths within a genome, which we hope may be helpful to ongoing and future annotation efforts. Due to their removal from transcripts prior to translation, intron sequences are generally not expected to respect the coding frame and meaning of the surrounding coding sequence. Correspondingly, many predicted introns in the most thoroughly annotated eukaryotic genomes have in frame stop codons, and predicted introns in these genomes are equally as likely to be a multiple of 3 basepairs (bp) ('3*n*'), and thus to conserve reading frame, as to contain an 'extra' one ($3n+1$) or two ($3n+2$) bp. For genomic sequences without exhaustive databases of transcript (EST and/or cDNA) sequences, prediction of introns is a difficult task. Here, examination of the distribution of intron lengths can provide insights into the possibility of intron over/underprediction.

We report distributions of predicted intron lengths for 29 fully sequenced eukaryotic species. We find frequent deviations in the number of predicted 3*n* introns relative to $3n+1$ and $3n+2$ introns. Some species show a pronounced deficit of 3*n* introns, others an excess of 3*n* introns. We discuss five different species that show highly skewed length distributions among predicted introns, and suggest ways in which current annotations might be improved.

*To whom correspondence should be addressed. Tel: +64-6-350-5515. Ext. 7626; Fax: +64-6-350-5626; Email: scottwroy@gmail.com

METHODS

We downloaded genome sequences and predicted gene sequences and coordinates as indicated in Table 1. For each *Entamoeba histolytica* gene with an annotated intron, we performed BLASTN searches of the corresponding genomic region (predicted intron plus 60 flanking upstream and downstream bases) against all *E. histolytica* reads in the NCBI Trace Archive and compared the assembled sequence against the best hit. As a negative control, we also BLASTed randomly selected uninterrupted predicted coding sequences of length 180 bp against the sequence reads. Only 132/5000 random sequence (2.6%) showed a gap, tenfold less than for the predicted intronic regions. In order to ensure that our observations were not due to errors in the GenBank files, we downloaded gene predictions from the individual websites for each genome showing deviations from equal proportions that is discussed below. In each case these annotations showed (nearly) identical proportions of introns in the three categories as found from the GenBank files. Novel perl scripts were written to perform the analyses described.

RESULTS AND DISCUSSION

Distribution of intron lengths across genome annotations for 29 species

We studied current genome sequence annotations for 29 different eukaryotic species (Table 2). For most genomes with large numbers of introns, there are very similar numbers of $3n+1$ and $3n+2$ introns: among genomes with >300 introns, the percentages of $3n+1$ and $3n+2$ introns are within 2.8% of each other in 23/25 genomes. In stark contrast, the number of $3n$ introns varies much more widely, falling only within 2.8% of the average of $3n+1$ and $3n+2$ for half (13/25) of the genomes. Species are roughly evenly divided between those that show an excess of $3n$ introns relative to $3n+1$ or $3n+2$ introns, and those that show a deficit of $3n$ introns. Two species, *E. histolytica* and the nucleomorph of *Bigelowiella natans*, show very different patterns—an excess of $3n+1$ introns in *B. natans* and an excess of $3n+2$ introns in *E. histolytica*. We next analyzed the sets of predicted introns for several cases that showed pronounced deviations from equal intron numbers in the three classes.

Excess of 3n introns: Thalassiosira pseudonana

Due to the lack of close relatives with sequenced genomes or very large samples of transcript sequences, genes in the *Thalassiosira pseudonana* genome were largely predicted by homology searches against sequences from deeply diverged eukaryotic species (8). Predicted *T. pseudonana* genes have on average 1.4 introns per gene. These predicted introns show a strongly skewed length distribution, with $3n$ introns accounting for 61.2% of all predicted introns (9573 $3n$ introns, 3037 $3n+1$, 3029 $3n+2$; an example is shown in Figure 1). Such a skew suggests that many predicted $3n$ introns are not true introns but instead represent exonic sequences. In keeping with this possibility, most $3n$ introns (75.2%) lack inframe stop codons, in stark contrast to $3n+1$ (29.1%) and $3n+2$ (28.6%) introns.

From these two observations it is possible to obtain independent estimates of the number of predicted introns that in fact represent coding sequence (i.e. false positive intron predictions). First, based on the assumption of equal numbers of $3n$, $3n+1$ and $3n+2$ introns there is a $3n$ excess of 6540 introns (note that numbers of $3n+1$ and $3n+2$ introns are nearly identical). Second, roughly equal fractions of $3n$, $3n+1$ and $3n+2$ introns are expected to lack inframe stop codons (for instance 29.1% of $3n+1$ introns and 28.6% of $3n+2$ introns should lack stop codons). There are 2368 stop-containing $3n$ introns, thus we expect roughly 972 (=2368 × 28.8%/[1–28.9%]) $3n$ introns without inframe stop codons, 6233 fewer than predicted. Thus, two independent estimates suggest that 6200–6600 (86–90%) of the 7205 predicted $3n$ stop codon-lacking introns instead represent unspliced coding sequence.

Deficit of 3n introns: Paramecium tetraurelia and Ostreococcus tauri

A second case suggestive of the reverse problem, of underprediction of $3n$ introns, is found in the annotation of the largest somatic chromosome of *Paramecium tetraurelia* (14). In this case, there is a striking deficit of predicted $3n$ introns (185 total) compared to $3n+1$ (436) and $3n+2$ (462) introns. In this case, this deficit is

Table 1. Distribution of lengths of predicted introns in complete genomes from 29 eukaryotic species

Species	Version	Source
Anopheles gambiae	AgamP3	Genbank
Apis mellifera	Amel2.0	Genbank
Arabidopsis thaliana	TAIR, version 5	Genbank
Aspergillus fumigatus	NC_007194.1-201.1	Genbank
Bigelowiella natans	DQ158856.1-8.1	Genbank
Caenorhabditis elegans	Wormbase 160	Wormbase
Ciona intestinalis	CINT1.95	Genbank
Cryptococcus neoformans	Version 1	Genbank
Cyanidoischyzon merolae	Version 1	Genbank
Dictyostelium discoideum	AAFI01000000.1	Genbank
Drosophila melanogaster	r4.3	Flybase
Encephalitozoon cuniculi	Version 1	Genbank
Entamoeba histolytica	AAFB01000000.1	Genbank
Fugu rubripes	FUGU4	EnsEMBL
Homo sapiens	NCBI35	EnsEMBL
Oryza sativa	REFSEQ	Genbank
Ostreococcus tauri	Oct 9 submission	Genbank
*Paramecium tetraurelia**	NC_006058.1	Genbank
Phytophthora sojae	Version 1.1	JGI
Phytophtora ramorum	Version 1.1	JGI
Plasmodium falciparum	3/10/2002 Version	PlasmoDb
Plasmodium yoelii	Version 1	PlasmoDb
Saccharomyces cerevisiae	NC_001133-48	Genbank
Schizosaccharomyces pombe	NC_00341.2-4.2	Genbank
Thalassiosira pseudonana	Thaps3	JGI
Toxoplasma gondii	ann3	TIGR
Trichomonas vaginalis	Version 1	Genbank
Ustilago maydis	Version 1	Genbank
Yarrowia lipolytica	CR382127.1-32.1	Genbank

likely due to the short intron lengths in *P. tetraurelia*: all predicted introns are less than 36 bp in length. Whereas, long non-coding sequences are likely to contain in-frame stop codons by chance, but short introns may lack stop codons, in which case $3n$ introns may be mistaken for coding sequence, whereas the presence of a $3n+1$ or $3n+2$ intron may be inferred from the disruption of the coding frame. That many stop codon-lacking $3n$ introns may have gone unpredicted is underscored by the high frequency of stop codons in the predicted $3n$ introns (91.3% contain a stop codon), much higher than in $3n+1$ (46.0%) or $3n+2$ (47.7%). If there were 264 $3n$ introns currently incorrectly predicted as coding sequence, the number of $3n$ introns would be equal to the average of $3n+1$ and $3n+2$, and the fraction of stop codon-containing introns would be similar (37.6% for $3n$) across classes.

A similar pattern is seen in the predictions for the *Ostreococcus tauri* genome, where $3n$ introns (20% of all predicted introns; 1290 total) are only half as frequent as

Table 2. Genome annotations for 29 fully sequenced species used in this study

Species	Introns	$3n$	$3n+1$	$3n+2$	Excess $3n$	$(3n+1)–(3n+2)$
Anopheles gambiae	37 901	0.405	0.298	0.297	0.108	0.001
Apis mellifera	145 454	0.376	0.309	0.316	0.063	−0.007
Arabidopsis thaliana	91 222	0.334	0.336	0.330	0.001	0.006
Aspergillus fumigatus	18 293	0.303	0.346	0.350	−0.045	−0.004
Aspergillus nidulans	24 772	0.319	0.342	0.339	−0.022	0.003
Bigelowiella natans	861	0.110	0.704	0.186	−0.334	0.518
Caenorhabditis elegans	137 752	0.332	0.335	0.333	−0.002	0.002
Ciona intestinalis	196 139	0.344	0.328	0.327	0.016	0.001
Cryptococcus neoformans	35 032	0.321	0.339	0.341	−0.019	−0.002
Cyanidoischyzon merolae	27	0.222	0.333	0.444	−0.167	−0.111
Dictyostelium discoideum	17 468	0.335	0.332	0.333	0.002	−0.001
Drosophila melanogaster	19 390	0.364	0.317	0.319	0.047	−0.002
Encephalitozoon cuniculi	15	0.267	0.467	0.267	−0.100	0.200
Entamoeba histolytica	3125	0.258	0.281	0.461	−0.113	−0.180
Fugu rubripes	171 912	0.314	0.340	0.346	−0.029	−0.006
Homo sapiens	307 019	0.330	0.333	0.337	−0.005	−0.003
Oryza sativa	100 262	0.341	0.330	0.329	0.012	0.002
Ostreococcus tauri	6450	0.200	0.402	0.398	−0.200	0.004
Paramecium tetraurelia	1082	0.170	0.403	0.427	−0.245	−0.024
Phanerochaete crysosporium	44 855	0.306	0.342	0.352	−0.041	−0.011
Phytophthora sojae	34 525	0.370	0.301	0.329	0.054	−0.028
Phytophtora ramorum	24 896	0.389	0.308	0.303	0.083	0.005
Plasmodium falciparum	7426	0.342	0.323	0.335	0.013	−0.012
Plasmodium yoelii	8143	0.347	0.321	0.333	0.020	−0.012
Saccharomyces cerevisiae	266	0.350	0.286	0.365	0.024	−0.079
Schizosaccharomyces pombe	4730	0.315	0.345	0.340	−0.028	0.005
Thalassiosira pseudonana	15 636	0.612	0.194	0.194	0.418	0.001
Toxoplasma gondii	27 495	0.336	0.331	0.332	0.004	−0.001
Trichomonas vaginalis	58	0.362	0.448	0.190	0.043	0.259
Ustilago maydis	4900	0.274	0.355	0.371	−0.088	−0.016
Yarrowia lipolytica	829	0.306	0.338	0.356	−0.040	−0.018

```
Intron 1

... GGA TTG GAA GGA CCT CCG GGC ACC gta agt ttt gtc tct tta ttg ttg gca ccg caa tta ctt
ttc tga gct gac aat taa cgc tgc tga aac tta aca tga ata tct ag* GGC AAA ACA ...

Intron 2

... GCA GCT CGC GTT Cgt aag tct ttg ttt cgc ctt gtc aaa cga acc acc ttc aat gct gac acg
ctg cgc cat atc tgt tca ttc ttc gat cac ata gGA AAC ACC TTC AGT CGA ...

Intron 3

... CGT GTA CAT GCG AAG AAg ttg ccg ggg ttt act gag tgt aag gga ata gac gag aag aga cct
gga agg tga gtt gtt gct gcg aag gga gct tgg caa gtt atg ctt tac gta ttg agc tga ctt ttt
act tta tag **C CTC GGG AAG GGA GCT AGT GTT GAT ...
```

Figure 1. Introns 1–3 and flanking sequences from *T. pseudonana* predicted gene 10 0621. Upper/lowercase sequence indicates predicted exonic/intronic sequence. Asterisks indicate frameshifts introduced by non-3n introns; intronic inframe stop codons are underlined. Intron 1 is an 86 bp intron (3n + 2) with two inframe stop codons. Intron 2 is an 84 bp intron (3n), which lacks inframe stop codons, and thus does not interrupt the ORF. Intron 3 is a 121 bp intron (3n + 1), which lacks inframe stop codons.

$3n+1$ (2592) or $3n+2$ (2567) introns (9). Again, a much higher fraction of predicted $3n$ introns contain stop codons (67.5%) than for $3n+1$ (42.1%) or $3n+2$ (38.5%). If there were an additional 1290 unpredicted $3n$ introns, there would be equal numbers across the classes, and similar fractions of stop-containing introns (33.8% for $3n$).

Difficulties associated with genome assemblies: *Entamoeba histolytica*

The above examples each concern difficulties of predicting introns based on an accurate genome sequence. Alternatively, errors in a genome assembly can lead to overprediction of introns. One example involves the genome of *E. histolytica* (10). Previous analysis of introns in genes thought to have been laterally transferred from prokaryotes showed that many predicted introns were associated with errors in the assembly in which a single base was missing in the assembly relative to the corresponding individual sequencing reads (11). Insertion of this missing base into the assembly yielded an ORF that continued through the predicted intronic sequence, suggesting that there is no intron present (e.g. Figure 2). Thus, these assembly indels led to frameshifts in coding sequences, which were compensated for by prediction of an intron.

Further analysis of the *E. histolytica* genome suggests that this may be a common problem. Among predicted *E. histolytica* introns, there is an excess of $3n+2$ introns (1449) over $3n$ (809) or $3n+1$ (878) introns. BLAST searches of the predicted intronic and flanking exonic sequences against individual sequencing reads showed that 23.1% (722/3126) of predicted introns were associated with gaps in the intron or within 120 bp of the intron (Figure 2). Of the gaps, 98.3% (831/845) were single-base gaps, and 81.3% (687/831) were missing bases in the assembly relative to the sequencing reads (the predominance of missing bases in the assembly is consistent with the excess of $3n+2$ introns, since the added base yields a $3n$ sequence; the smaller number of extra assembly bases relative to reads is consistent with the smaller excess of $3n+1$ introns over $3n$ introns).

Correction of these apparent assembly errors extended the ORF through the intronic sequence in 79.8% (538/649) of these cases, suggesting that the predicted intronic sequence instead represents coding sequence. In an additional 48 cases (7.4%), correction of the apparent assembly error yields an ORF spanning from within (or upstream of) the predicted intron to the predicted stop codon (or to the next predicted intron boundary in the case of multi-intron genes). These results suggest that at least some 20% of predicted *E. histolytica* introns are not in fact introns but instead coding sequence. Thus, upwards of 6% of predicted genes in the *E. histolytica* genome appear to have an assembly error within their sequence.

The peculiar case of the *B. natans* nucleomorph

Finally, it is worth noting that there is at least one apparently bona fide case of stiking genome-wide difference between introns of different length classes (12,13).

```
... CTT TTT AAT TCA TTC ATT TTA TTT AAA AAA gtt tca cta ttt ttt tag TTA TTA ATC AAA ATT
TTT AAA CTT AGA TTA TAT TTT TAT AAT AAA CCT TGT TCT TCA ACT AAA TAT ACA CCA GAG AAA TGT
AAA ATA TTG GAA Agt act tta taa ttt acc ag   AA AAA TAT GCT AAT TTA ATA GTA AAA AAA GTT
TTA TTG GAT AAA GTA TTT ATT AAA gta atc tat aaa taa tag AAT CTC ATC AAG AAT TTC AAC CAT
AAT AAT ATT CAT AAA CAA CAA CAC GAC ATC AAA G ...
```

Figure 2. *Bigelowiella natans gene ABA27371.1*, introns 1–3 and flanking sequence. Note that introns 1 and 3 are 18 bp, but also note that they contain stops, and therefore disrupt the ORF. As discussed in the text, other 18 bp introns without stops may have excaped prediction.

```
A
     L   V   P   R   L   T   P   I   L   R   N   T   N   E   T   I   E
... CTT GTA CCT CGT TTA ACT CCA ATA TTA AGA AAT ACT AAT GAA ACA ATT GAA

 E   A   C   I   G   L   I   G   I   I   A   K   K   S   A   D   T   G
GAA GCA Tgt att gga tta att gga att att gca aag aaa tca gCT GAT ACA GGT

 A   E   M   V   H   L   K   E   W   M   R   I   C   H   E   L   L   D   A
GCT GAA ATG GTT CAT TTA AAA GAA TGG ATG AGA ATT TGT CAT GAA TTA CTT GAT GCA ...


B
     E   Y   H   A   R   V   A   E   R   H   T   G   S   G   I   D   P   T   Q   I   V
... GAA TAT CAT GCA CGT GTT GCT GAA AGA CAT ACT GGA TCT GGA ATT GAT CCA ACC CAA ATT gtt

 E   Q   H   L   P   L   P   T   D   R   I   L   L   P   K   T   I   E   E   K   L   L
gaa caa cat ctt cct tta cct aca gat agA ATA TTA TTA CCT AAA ACA ATA GAA GAG AAA TTA TTA

 C   S   A   D   K   F   F   S   K   S
TGT TCT GCT GAT AAG TTC TTC TCA AAA TCA
```

Figure 3. Assembly indels in *E. histolytica* lead to artifactual intron predictions. Predicted intron (lowercase) and flanking exon (uppercase) sequence is shown. The underlined base is present in sequencing reads but absent in the assembly. Bold amino acid sequencs are unique to the corrected sequence. (**A**) Gene EAL42479.1. An assembly deletion of a single A lead to prediction of a 35 bp intron (3n + 2) spanning the deletion position to correct the coding frame. (**B**) Gene EAL42572. An assembly deletion of a single G lead to prediction of a 32 bp intron (3n + 2) downstream of the deletion.

Among predicted introns in the genome of the nucleomorph of the *B. natans* genome, 70.3% of introns are $3n+1$ (12,13). However, this is due to the extreme regularity of intron length, with 70.2% of predicted introns having length 19 bp, and 99.1% being of length 18–20 bp (an example is given in Figure 3).

How accurate is this length distribution likely to be? Introns in *B. natans* are predicted based on maximizing ORF length. On its face, then, we would expect the rates of detection of 19 and 20 bp introns to be similar (since both impose a frameshift). Thus, the great excess of 19 bp introns over 20 bp introns is likely to be a true feature of intron lengths in *B. natans*. On the other hand, detection of 18 bp introns will be far more difficult since these introns will not disrupt the ORF unless they contain an inframe stop codon (which is not likely in a 18 bp sequence). Correspondingly, the fraction of predicted introns containing inframe stop codons is higher for 18 bp introns (89.7%) than for 19 bp introns (48.4%). This suggests that 18 bp introns are underpredicted by perhaps twofold, however underprediction is unlikely to account for the 7.0-fold excess of predicted 18 bp introns over 19 bp introns.

Integrating genome assembly and annotation

The example of *E. histolytica* raises a larger point of the possible utility of integrating genome assembly with gene annotation. In this case, examination of predicted genes indicated errors in the genome assembly itself, thus integration of the two processes should lead to improvements in both assembly and annotation quality. Such considerations are likely to be all the more important with the increased number of partial and low-coverage genome sequencing projects. Analysis of the apparent coding meaning of preliminary assemblies could identify probable indels in the assembly, and corresponding individual sequencing reads could then be scrutinized in order to correct actual errors. This process lends itself well to automation and we think that such feedback between assembly and annotation could potentially substantially diminish assembly errors in coding regions.

Exploiting intron length distributions in genome annotation pipelines

In the absence of extensive transcript data, intron prediction often proceeds by statistical comparison of alternative gene models: introns are called when an intron-containing structure has a higher probability given model perameters than does an intron-lacking structure (3). In the case of non-$3n$ or stop-containing introns, this is comparatively straightforward, with intron prediction being likely if the resulting coding sequence is significantly longer than the intronless alternative. For non-stop-containing $3n$ introns, the case is quite different. In this case intron calling involves comparison of two structures with identical 5′ and 3′ flanking coding sequences, thus whether an intron is called will depend only on the extent to which the intervening sequence conforms to expected intron sequences. Here, too much/little sensitivity to intron-like structures will lead to over/under-prediction of introns.

One way to address such potential problems would be to introduce an additional training step into gene annotation. Intron sensitivity (i.e. prior probability of intron calling) could be automaticallly fine tuned based on a training set (genomic region) with unknown intron coordinates until predicted intron distributions (both in terms of lengths and in terms of stop frequencies within introns from different categories) became similar to expected values. Such a procedure is likely to be very helpful in species for which very few verified gene structures are known, since in such cases sensitivity in intron calling is otherwise difficult to gauge. Trial and error will no doubt be required to arrive at a fully functioning protocol, however we suspect that such efforts would be well rewarded by increased accuracy of intron and therefore proteome prediction.

Alternative splicing and intron length distribution

One possible actual biological deviation from equal proportions is worthy of discussion. In a subset of alternative splicing events, an intronic sequence from one splicing isoform is retained in another sequence (so-called ‘intron retention’). In such cases, the alternatively spliced intron must be $3n$ and lack inframe stop codons in order for both isoforms to encode proteins. Thus, frequent alternative splicing could conceivably bias intron lengths towards $3n$ introns.

Four observations suggest that such alternative splicing events are not a major contributor to the skewed distributions reported here. First, the genomes which show the highest frequencies of alternative splicing (for instance *Drosophila melanogaster, Caenorhabditis elegans, Homo sapiens and Arabidopsis thaliana*) are among the genomes that show the intron length proportions most nearly equal. Second, in all studied genomes, not more than 5% of introns are known to be alternatively spliced in the manner discussed above, and many of these are not $3n$ introns, thus this effect is unlikely to drive any more than a small bias even in the most highly alternatively spliced genomes. Third, there is no evidence for frequent alternative splicing in any of the genomes that show pronounced intron length skew. Finally, such an effect specifically predicts an excess of $3n$ introns, and as such is not a contributor to the other observed skews ($3n$ deficit, $3n+2$ excess).

Concluding remarks

Accurate genome annotation is an extremely difficult problem, requiring balancing of false negatives and positives, and accuracy versus time constraints. Even the best annotation sets are subject to improvement. Evaluation of distributions of predicted intron lengths promises rapid and straightforward detection of a variety of possible systematic biases in gene prediction or even, as in the case of *E. histolytica*, problems with genome assemblies.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Allan Wilson Centre.

Conflict of interest statement. None declared.

REFERENCES

1. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.*, **24**, 3439–3452.
2. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
3. Gelfand,M.S. and Roytberg,M.A. (1996) Prediction of the exon-intron structure by a dynamic programming approach. *Biosystems*, **30**, 173–182.
4. Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, **3**, 698–709.
5. Pavlovic,V., Garg,A. and Kasif,S. (2002) A Bayesian framework for combining gene predictions. *Bioinformatics*, **18**, 19–27.
6. Allen,J.E., Pertea,M. and Salzberg,S.L. (2004) Computational gene prediction using multiple sources of evidence. *Genome Res.*, **14**, 142–148.
7. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.
8. Armbrust,E.V., Berges,J.A., Bowler,C., Green,B.R., Martinez,D., Putnam,N.H., Zhou,S., Allen,A.E., Apt,K.E. *et al.* (2004) The genome of the diatom *Thalassiosira pseuodonana*: ecology, evolution, and metabolism. *Science*, **306**, 79–86.
9. Derelle,E., Ferraz,C., Rombaut,S., Rouze,P., Worden,A.Z., Robbens,S., Partensky,F., Degroeve,S., Echeynie,S. *et al.* (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl Acad. Sci. USA*, **103**, 10647–10652.
10. Loftus,B., Anderson,I., Davies,R., Alsmark,U.C., Samuelson,J., Amedeo,P., Roncaglia,P., Berriman,M., Hirt,R.P. *et al.* (2005) The genome of the protist parasite *Entamoeba histolytica*. *Nature*, **433**, 865–868.
11. Roy,S.W., Irimia,M. and Penny,D. (2006) Very little intron gain in *Entamoeba histolytica* genes laterally transferred from prokaryotes. *Mol. Biol. Evol.*, **23**, 1824–1827.
12. Gilson,P.R. and McFadden,G.I. (1996) The miniaturized nuclear genome of eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc. Natl Acad. Sci. USA*, **93**, 7737–7742.
13. Gilson,P.R., Su,V., Slamovits,C.H., Reith,M.E., Keeling,P.J. and McFaddden,G.I. (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl Acad. Sci. USA*, **103**, 9566–9571.
14. Zagulski,M., Nowak,J.K., Le Mouël,A. , Nowacki,M., Migdalski,A., Gromadka,R., Noël,B., Blanc,I., Dessen,P., *et al.* (2004) High coding density on the largest *Paramecium*.