Cleaning the GenBank *Arabidopsis thaliana* data set

Peter G. Korning, Stefan M. Hebsgaard, Pierre Rouzé[1] and Søren Brunak*

Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Lyngby, Denmark and [1]Laboratoire Associé de l'INRA, VIB, University of Ghent, B-9000 Gent, Belgium

Received September 15, 1995; Revised and Accepted November 27, 1995

ABSTRACT

Data driven computational biology relies on the large quantities of genomic data stored in international sequence data banks. However, the possibilities are drastically impaired if the stored data is unreliable. During a project aiming to predict splice sites in the dicot *Arabidopsis thaliana*, we extracted a data set from the *A.thaliana* entries in GenBank. A number of simple 'sanity' checks, based on the nature of the data, revealed an alarmingly high error rate. More than 15% of the most important entries extracted did contain erroneous information. In addition, a number of entries had directly conflicting assignments of exons and introns, not stemming from alternative splicing. In a few cases the errors are due to mere typographical misprints, which may be corrected by comparison to the original papers, but errors caused by wrong assignments of splice sites from experimental data are the most common. It is proposed that the level of error correction should be increased and that gene structure sanity checks should be incorporated—also at the submitter level—to avoid or reduce the problem in the future. A non-redundant and error corrected subset of the data for *A.thaliana* is made available through anonymous FTP.

INTRODUCTION

Biological sequence databases offer scientific researchers unique opportunities for working with large quantities of genomic data. The GenBank, EMBL and DDBJ databases aim to contain all published DNA sequences. In its feature table each entry holds information about transcription, splicing and translation associated signals. This information may be used to create large data subsets, where the sequences are related to their functionality.

Unfortunately, GenBank is currently not optimally suited for such extraction. Too often errors occur in the entries. Almost all major publicly available genomic databases suffer from this condition. The problem of corrupt databanks has been pointed out earlier (1,2). However, the situation remains more or less the same.

For the construction of a data set relating the DNA (or pre-mRNA) sequence to its alternating exon–intron structure, we extracted relevant information from the *Arabidopsis thaliana* entries in GenBank ver. 87 (3). The data set has been used in work aiming to predict splice sites in the dicot by means of neural network algorithms (Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P. and Brunak, S., submitted) and (4). A common feature of neural network algorithms is their ability to cope with non-linearities in the association between objects and categories. During training on the initially extracted data training difficulties were encountered for many of the splice sites (1,2). This led us to perform a range of 'sanity' checks that revealed quite a large percentage of errors in the *A.thaliana* entries. The most effective of these checks being examination of particular splice sites, which diverged strongly from the 'extended consensus sequence' of the organism. The latter was created from a measurement of the Shannon information around the splice sites and will be described below. Of the 999 *A.thaliana* entries in GenBank rel. 87, 167 contained enough splicing information to be of interest to our work.

A list of corrupt *A.thaliana* GenBank entries has been compiled and in most cases we explain how to correct the errors. When possible, the original contributers of the entries have been contacted and the errors should be corrected in future versions of GenBank. The errors detected are often shift errors, which are not caused by misprints only. Instead, many errors are created when splice sites are assigned by homology with wrong cDNA assignments. We suggest that the consistency of the annotation is assessed prior to the submission in order to reduce the error rate. Such proof-reading may be assisted by using internet available servers, which are able to evaluate the correlation between the nucleotides in a single splice site from a consensus view point, rather than relying only on a more individual assignment by hand or by cDNA homology.

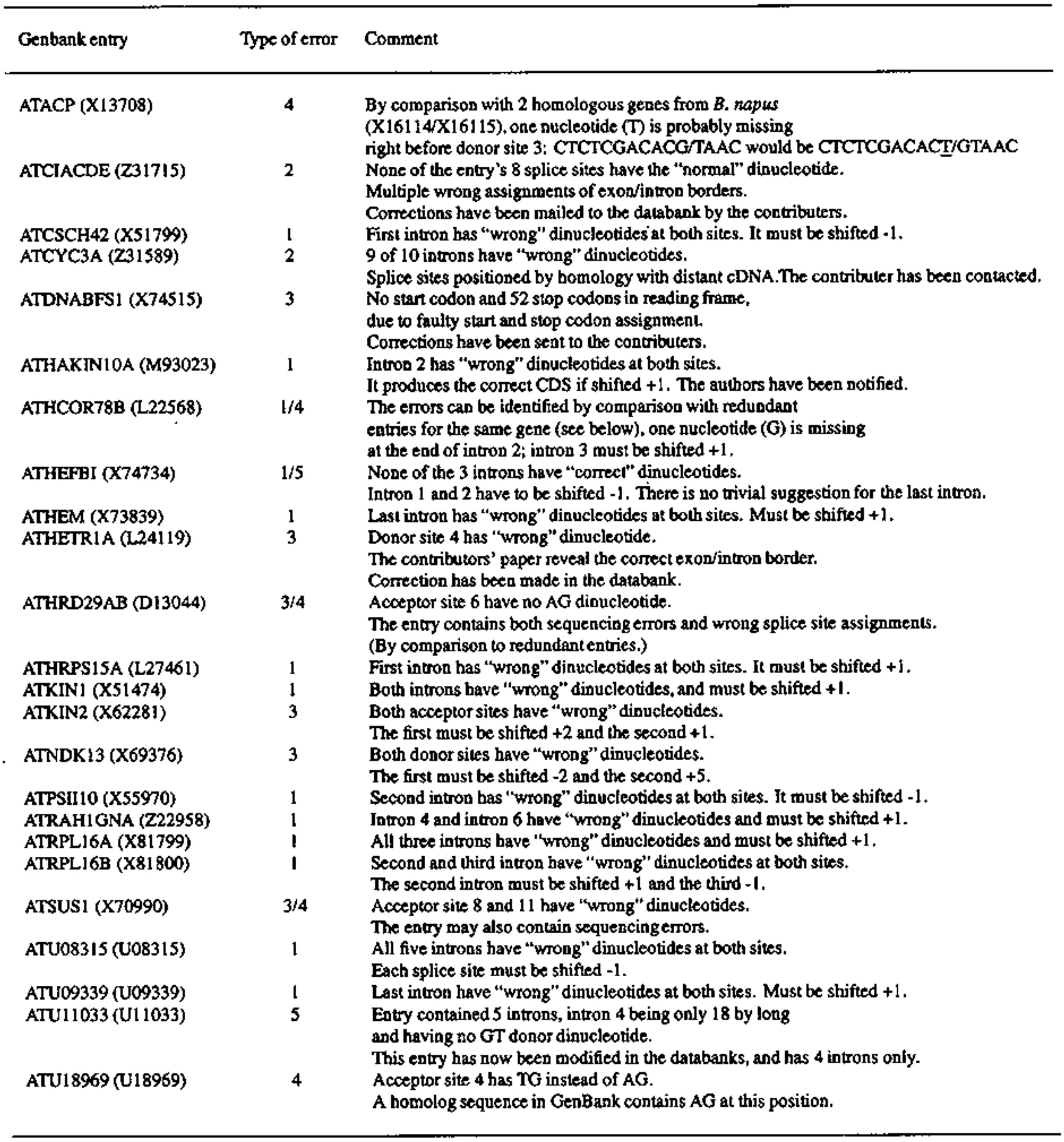
MATERIALS AND METHODS

An initial data set was extracted from GenBank 87.0 by use of a software scanner. All *A.thaliana* entries were examined. The criteria for inclusion of an entry in the data set were the following: (i) it must contain two or more introns; (b) it must not in any way be indicated as being partial; and (iii) there must be no logical conflicts between the entry's description of its components, e.g. a substring which is defined as being both an intron and an exon. (i) Means that all genes in the data set contain internal exons, while (ii) means that the CDS in a given entry in the data set is neither mentioned as being partial in the DEFINITION, nor as being partial or as having uncertain borders in the FEATURES section. Due to the third criterion 20 entries were excluded, having direct annotation conflicts not caused by alternative splicing.

After this reduction 167 entries remained in the data set and training of the neural network algorithms following (5) was

[*] To whom correspondence should be addressed

Table 1. The *A.thaliana* entries found in GenBank rel.87 likely to be erroneous



Genbank entry	Type of error	Comment
ATACP (X13708)	4	By comparison with 2 homologous genes from *B. napus* (X16114/X16115), one nucleotide (T) is probably missing right before donor site 3: CTCTCGACACG/TAAC would be CTCTCGACACT/GTAAC
ATCIACDE (Z31715)	2	None of the entry's 8 splice sites have the "normal" dinucleotide. Multiple wrong assignments of exon/intron borders. Corrections have been mailed to the databank by the contributers.
ATCSCH42 (X51799)	1	First intron has "wrong" dinucleotides at both sites. It must be shifted -1.
ATCYC3A (Z31589)	2	9 of 10 introns have "wrong" dinucleotides. Splice sites positioned by homology with distant cDNA.The contributer has been contacted.
ATDNABFS1 (X74515)	3	No start codon and 52 stop codons in reading frame, due to faulty start and stop codon assignment. Corrections have been sent to the contributers.
ATHAKIN10A (M93023)	1	Intron 2 has "wrong" dinucleotides at both sites. It produces the correct CDS if shifted +1. The authors have been notified.
ATHCOR78B (L22568)	1/4	The errors can be identified by comparison with redundant entries for the same gene (see below), one nucleotide (G) is missing at the end of intron 2; intron 3 must be shifted +1.
ATHEFBI (X74734)	1/5	None of the 3 introns have "correct" dinucleotides. Intron 1 and 2 have to be shifted -1. There is no trivial suggestion for the last intron.
ATHEM (X73839)	1	Last intron has "wrong" dinucleotides at both sites. Must be shifted +1.
ATHETR1A (L24119)	3	Donor site 4 has "wrong" dinucleotide. The contributors' paper reveal the correct exon/intron border. Correction has been made in the databank.
ATHRD29AB (D13044)	3/4	Acceptor site 6 have no AG dinucleotide. The entry contains both sequencing errors and wrong splice site assignments. (By comparison to redundant entries.)
ATHRPS15A (L27461)	1	First intron has "wrong" dinucleotides at both sites. It must be shifted +1.
ATKIN1 (X51474)	1	Both introns have "wrong" dinucleotides, and must be shifted +1.
ATKIN2 (X62281)	3	Both acceptor sites have "wrong" dinucleotides. The first must be shifted +2 and the second +1.
ATNDK13 (X69376)	3	Both donor sites have "wrong" dinucleotides. The first must be shifted -2 and the second +5.
ATPSII10 (X55970)	1	Second intron has "wrong" dinucleotides at both sites. It must be shifted -1.
ATRAH1GNA (Z22958)	1	Intron 4 and intron 6 have "wrong" dinucleotides and must be shifted +1.
ATRPL16A (X81799)	1	All three introns have "wrong" dinucleotides and must be shifted +1.
ATRPL16B (X81800)	1	Second and third intron have "wrong" dinucleotides at both sites. The second intron must be shifted +1 and the third -1.
ATSUS1 (X70990)	3/4	Acceptor site 8 and 11 have "wrong" dinucleotides. The entry may also contain sequencing errors.
ATU08315 (U08315)	1	All five introns have "wrong" dinucleotides at both sites. Each splice site must be shifted -1.
ATU09339 (U09339)	1	Last intron have "wrong" dinucleotides at both sites. Must be shifted +1.
ATU11033 (U11033)	5	Entry contained 5 introns, intron 4 being only 18 bp long and having no GT donor dinucleotide. This entry has now been modified in the databanks, and has 4 introns only.
ATU18969 (U18969)	4	Acceptor site 4 has TG instead of AG. A homolog sequence in GenBank contains AG at this position.

Type of error indicates: 1, equal shift of donor and acceptor sites (typically one or two nucleotides); 2, unequal shift of donor and acceptor sites; 3, mispositioning of one splice site, or other misplacing of a feature; 4, sequencing errror suggested; 5, other case/unknown.

performed. The output neuron of the networks contained one unit, trained to classify the central nucleotide in the DNA segment as being either a splice site or a non-splice site. It soon became clear that unless very large networks were used, certain sites in the data set could just not be learned. This phenomenon is symptomatic of impurities in the input data. As a result of this, a number of sanity checks was performed on the data set. These included: checks for reading frame inconsistencies, checks for presence/lack of start and stop codons, checks for introns below the minimal functional intron length, and finally investigation of splice sites deviating significantly from their extended consensus sequence.

The extended consensus sequence

The general sequence patterns were found by plotting the Shannon information content in the context of aligned splice sites in the data set (6). The formula used was

$$H(i) = -\sum_{\alpha=1}^{4} P_i^{\alpha} \log_2 P_i^{\alpha} \qquad \mathbf{1}$$

where $H(i)$ is the Shannon information at position i, and P^{α}_i is the probability of finding nucleotide α (α {A, C, G, T}) at position i. The probabilities were computed from the frequencies of the nucleotides in the data set. Plotted together with the nucleotide frequencies at each locus, such curves can be said to constitute an extended consensus sequence of the splice sites. The most conserved nucleotides in the *A.thaliana* consensus sequences found were AG|GTAAGT and TGYAG|GT for donor and acceptor sites respectively.

RESULTS

Reading frame inconsistencies

First, in each gene it was examined whether the sum of the number of nucleotides in the translated parts of the exons did obey the modulo three rule and thus did not contain incomplete codons. There turned out to be one entry that did not obey the modulo three rule, namely ATDNABFS1.

Start and stop codons

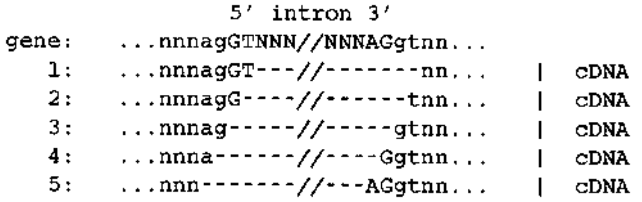
Secondly, a scan for the lack of start codons (ATG), for the lack of stop codons [TAA TAG and TGA, *A.thaliana* uses the standard genetic code (ftp://weeds.mgh.harvard.edu/pub/codon/ath.cod)], and for the presence of stop codons inside the reading frame was performed. The entry ATDNABFS1 turned out to lack a start codon and contained no less than 52 stop codons in the reading frame. Correspondence with the author revealed that a fair CDS is found when the first nucleotide of the start codon is taken as position 393 instead of position 374, and the first nucleotide of the stop codon is taken as position 3134 instead of position 3305.

Minimal functional intron length

Thirdly, the data set was scanned for entries containing very short introns. As they have to form a lariat, splicing becomes functionally impossible if the intron is very short. According to a recent review (7,8) a minimal functional length of 64 nt is reported for dicots. While it seems that this number should be lowered somewhat (our scan revealed seemingly correct introns of lengths between 58 and 64 nt), an intron of length 18 nt as intron four in ATU11033 was an error (has later been corrected by the authors). This entry was removed from the data set.

Strange non-consensus splice sites

The largest number of errors was found by examining apparent deviations from the donor and acceptor splice site consensus sequences. Of the 878 donor sites and 880 acceptor sites in the extracted data set, 45 donor sites and 44 acceptor sites turned out to lack the normal dinucleotides, GT and AG. The non-consensus splice sites were found in 26 entries only, which were examined manually by comparing to the original and related publications, and by searching for related entries in the databases. From this careful examination we strongly suggest that *all of the acceptor sites and all the donor sites but three are errors*! The three donor sites represent true exceptions. They appear in the entries ATHACOACAR (donor site 15), ATHFUS6A (donor site 3) and ATHHANKA (donor site 6). The use of GC instead of GT at the donor site is observed in the order of ~1% of the *A.thaliana* introns. Their occurrence as true donor sites has recently been reported in the myrosinase genes from *Brassicaceae*, the family including the species *A.thaliana* (9).



```
                  5' intron 3'
gene:    ...nnnagGTNNN//NNNAGgtnn...
   1:    ...nnnagGT---//-------nn...     |  cDNA
   2:    ...nnnagG----//------tnn...     |  cDNA
   3:    ...nnnag-----//-----gtnn...     |  cDNA
   4:    ...nnna------//----Ggtnn...     |  cDNA
   5:    ...nnn-------//---AGgtnn...     |  cDNA
```

Figure 1. Alignments of a genomic sequence to its cognate cDNA. The top line represents the sequence of a gene around an intron, showing its 5′ and 3′ borders, while the next five lines represent alternative alignments of the cDNA sequence. Only alignment 3 gives the correct splice site assignment, leaving out of the alignment the proper intron sequence starting by ‘GT’ and ending by ‘AG’ dinucleotides.

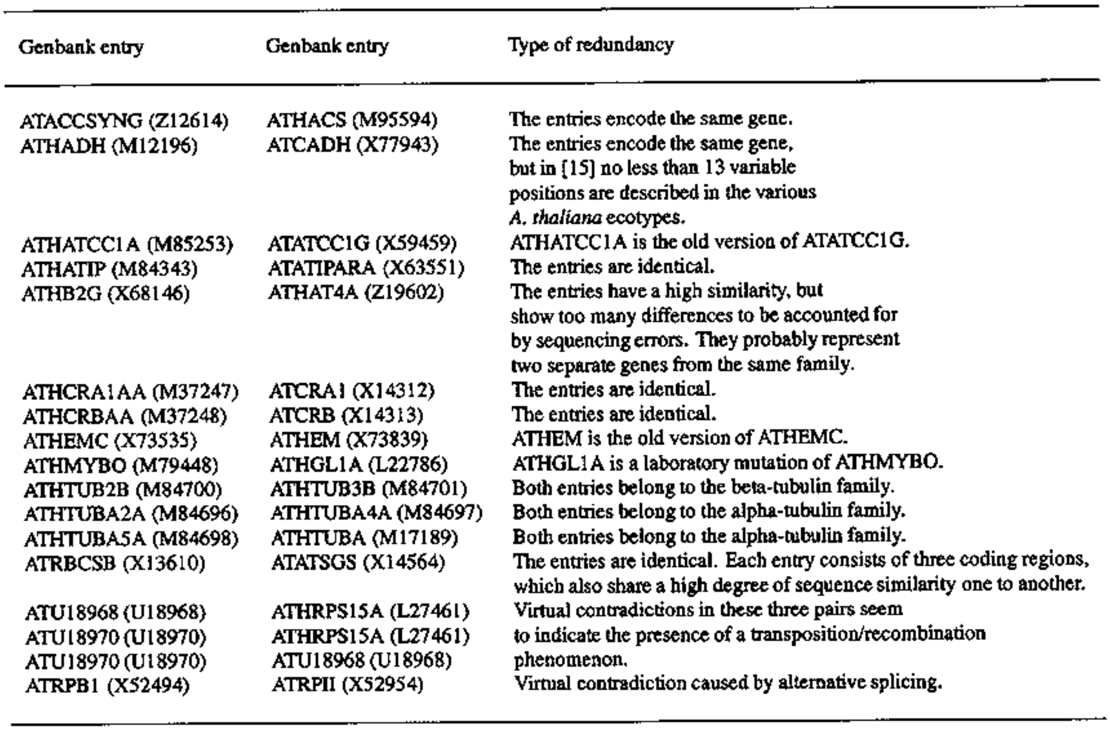
A list of the remaining entries, with comments on the errors, can be found in Table 1. The errors may be divided into a small number of categories. A large part of them (21/24) are caused by bad localization of a feature in the sequence, most often the splice sites. If the assignment is performed by comparing the gene sequence and homologous cDNA without caring for a proper consensus, this kind of mistake may occur (see Fig. 1). Since the sequences around the 5′ and 3′ splice sites often are similar, alignment ambiguities will result and one among several good solutions may be discarded. For this kind of errors, anyone has access to the same basic information as the authors were having themselves and sometimes even more due to new data and publications that have appeared later. In a few cases other kinds of errors are indicated. Their occurence may be related to sequence errors coming from editing or from the sequencing itself. Here the authors only may be able to track down and confirm the suggested errors, albeit the indication of errors is always very strong. Examples of such indications are given by redundant entries for the same gene (ATHRD29AB), or by entries for very close homologs (ATACP, ATU18969), or by clear biological inconsistencies (ATU11033). This latter case is among those where entries have been modified in a later version of GenBank, confirming our suggestion of error. In several cases we have been able to contact the authors, ending up with consistent modifications of the entry for most of the genes.

The shifts on each side are sometimes unequal (e.g. –1, +2), but nevertheless keep the coding sequence in frame. This concerns cases where splice site assignments have been done by the authors using a distant cDNA homolog and ignoring the consensus at the borders. Only tentative assignments can be suggested in such a situation, a secure one awaiting better homologs. There are two entries of that kind (ATCIADE, ATCYC3A) for which the authors have been contacted, resulting in a modified entry for one of them. However, in most cases the shifts needed to produce a good fit to the consensus, and keeping the frame with a similar (most of the time identical) coding sequence, are equal for one or more introns from 13 out of the 24 incorrect entries. This common kind of error is not a consequence of bad database management or misprints, but stems from the experimentalists who submitted the sequence.

Redundant sequences

It is well known that the data in GenBank have a certain degree of redundancy. Therefore the length of the longest common substring

Table 2. The redundant *A.thaliana* entries with comments on the redundancy type



Genbank entry	Genbank entry	Type of redundancy
ATACCSYNG (Z12614)	ATHACS (M95594)	The entries encode the same gene.
ATHADH (M12196)	ATCADH (X77943)	The entries encode the same gene, but in [15] no less than 13 variable positions are described in the various *A. thaliana* ecotypes.
ATHATCC1A (M85253)	ATATCC1G (X59459)	ATHATCC1A is the old version of ATATCC1G.
ATHATIP (M84343)	ATATIPARA (X63551)	The entries are identical.
ATHB2G (X68146)	ATHAT4A (Z19602)	The entries have a high similarity, but show too many differences to be accounted for by sequencing errors. They probably represent two separate genes from the same family.
ATHCRA1AA (M37247)	ATCRA1 (X14312)	The entries are identical.
ATHCRBAA (M37248)	ATCRB (X14313)	The entries are identical.
ATHEMC (X73535)	ATHEM (X73839)	ATHEM is the old version of ATHEMC.
ATHMYBO (M79448)	ATHGL1A (L22786)	ATHGL1A is a laboratory mutation of ATHMYBO.
ATHTUB2B (M84700)	ATHTUB3B (M84701)	Both entries belong to the beta-tubulin family.
ATHTUBA2A (M84696)	ATHTUBA4A (M84697)	Both entries belong to the alpha-tubulin family.
ATHTUBA5A (M84698)	ATHTUBA (M17189)	Both entries belong to the alpha-tubulin family.
ATRBCSB (X13610)	ATATSGS (X14564)	The entries are identical. Each entry consists of three coding regions, which also share a high degree of sequence similarity one to another.
ATU18968 (U18968)	ATHRPS15A (L27461)	Virtual contradictions in these three pairs seem to indicate the presence of a transposition/recombination phenomenon.
ATU18970 (U18970)	ATHRPS15A (L27461)	
ATU18970 (U18970)	ATU18968 (U18968)	
ATRPB1 (X52494)	ATRPII (X52954)	Virtual contradiction caused by alternative splicing.

which would occur in a collection of random strings corresponding in size to the data set was calculated. Then a scan for common substrings above this length was performed and entries where such substrings did occur were investigated manually. There are basically two classes of redundancies: First, there are cases where the same gene has been submitted twice; either by the same author or by different authors. These sequences tend to vary a bit in length, but otherwise be more or less identical. Secondly, there are cases where two different genes are members of the same gene family and so closely related that only one of them should be included in the data set (see Table 2 for the redundant sequences and comments).

The redundancy reduction has been made in order not to overestimate the performance of intron splice site finding computer programs. If the aim was, for example, to study alternative splicing this would be another matter. For quite some time a rigorous procedure for this removal has been common practice in the field of protein structure prediction (10–12). We are in the process of determining a proper alignment threshold for nucleotide data making it possible to distinguish between cases where the splice sites may be found safely by alignment, and those where genuine prediction is needed (Tolstrup, N., Dalsgaard, K., Engelbrecht, J. and Brunak, S., in preparation.) Today, algorithms are tested on an ill-defined mixture of the two.

Virtual contradictions

A search for ‘virtual contradictions’ in the data set was also performed. A virtual contradiction for a given substring-size is defined as two identical sequences, where the central nucleotides have different functionalities, e.g. splice site and non-splice site, or coding nucleotide and non-coding nucleotide. Four such pairs of virtual contradictions were found (see Table 2). The contradiction between the pair ATRPBl and ATRPII seems to be caused by alternative splicing. More interestingly, however, are the contradictions between the three pairs ATU18968/ATHRPSlSA, ATU18970/ATHRPS15A and ATU18970/ATU18968 respectively. Although further analysis is needed and actually undertaken (Rouzé, P., Breyne, P. and Van Gysel, A., in preparation), these virtual contradictions may stem from a new kind of recombination and/or transposition event. Interestingly enough, this is not mentioned by the contributors in the corresponding papers (13,14).

DISCUSSION

It is clear that the errors found may be divided into a few categories. Many of the errors are simple shift errors, created somewhere along the path from the laboratory to the database, which can be easily corrected by comparison of the sequence to its corresponding cDNA if the latter is available (taking into account well documented consensus sequences). Other errors are caused by sequencing or typing errors. There is however a class of errors, which are more or less impossible to correct, unless one wants to go back to the laboratory. These errors occur when a DNA string, having been sequenced correctly, is assigned splice sites by homology with a more or less distant species. It would be highly desirable if such speculative assignments were stated clearly in GenBank, as the quality of such entries is often not good enough for creating powerful data driven computer based methods. Splice sites shifted a few nucleotides, constitute directly ‘negative information’ to for example neural network algorithms, and are extremely harmful to their learning and generalisation capabilities if the errors are frequent.

After the above mentioned erroneous entries had been corrected or removed, very clear consensus sequences did emerge from the Shannon plots. They correspond to the ones mentioned above, but with much more certainty at each position. The contrast to the Splice

Site Consensus Table in AAtDB (ftp://weeds.mgh.harvard.edu/aatdb-info/Splice_Site_Consensus) is striking: in AAtDB a more disorderly picture emerges; >6% (maybe up to 12%, only single nucleotide frequencies are given) of the donor sites do not contain the GT dinucleotide. The situation is the same for AG at the acceptor sites. The *A.thaliana* sequences from AAtDB (rel. 3–5) may have been selected by other criteria than the ones adopted in this work. Still, observing differences as large as these, we do believe that AAtDB might benefit from a thorough sanity scan as well.

One reason for the success of finding the above mentioned errors in GenBank, is the fact that complete coding sequences were used only. These have to obey a large number of rules in order to be processed correctly by the cell's machinery. However, the consensus sequence check can be used anywhere splicing occurs.

Biological research does now, and will even more in the future, depend on information retrieved from large genomic databases. As they control the information flow between scientists, it would be an obvious place for proof-reading of the data, at least by computers and preferably checked by human experts. The examples from this paper demonstrate the necessity of such proof-reading. From similar work on human genes, and on genes from *C.elegans*, we estimate that the GenBank error rate is somewhat smaller for boths organisms, in the order of 5–10%. The error rate is likely to depend of the average amount of work done on the entries, but also the diversity of the splice sites, from a given organism.

DATA SET AND PREDICTION SERVER AVAILABLE

The cleaned *A.thaliana* data set is made available to other researchers by anonymous FTP from ftp.cbs.dtu.dk:/pub/arabidopsis/NetPlantGene.seq. The current version contains 147 genes from 145 GenBank entries all in all. The intron splice site prediction method has also been made available by electronic mail and through the WWW (Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S., submitted). Send a file containing the word 'help' to the internet address NetPlantGene@cbs.dtu.dk to obtain information on sequence formats and other details.

ACKNOWLEDGEMENTS

The authors wish to thank Anders Gorm Pedersen for helpful discussion, and Krzysztof Rapacki for providing the primary GenBank scanner. This work was supported by the Danish National Research Foundation.

REFERENCES

1 Brunak, S., Engelbrecht, J. and Knudsen, S. (1990) *Nature* **343**, 123.
2 Brunak, S., Engelbrecht, J. and Knudsen, S. (1990) *Nucleic Acids Res.* **18**, 4797–4801.
3 Meyerowitz, E.M. and Somerville, C.R. (eds) (1994) *Arabidopsis*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
4 Hertz, J., Krogh, A. and Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation*, Addison-Wesley.
5 Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) *J. Mol. Biol.* **220**, 49–65.
6 Shannon, C.E. (1948) *Bell System Tech. J.* **27**, 379–423, 623–656.
7 Filipowicz, W., Gniadkowski, M., Klahre, U. and Liu, H.-X. (1995) In Lamond, A.I. (ed.) *Pre-rnRNA Processing, Pre-mRNA splicing in plants*, R.G. Landes Company, pp. 65–77.
8 Goodall, G.J. and Filipowicz, W. (1990) *Plant Mol. Biol.* **14**, 727–733.
9 Xue, J. and Rask, L. (1995) *Plant Mol. Biol.* **29**, 167–171.
10 Sander, C. and Schneider, R. (1991) *Proteins* **9**, 56–68.
11 Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Sci.* **1**, 409–417.
12 Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. (1995) *Proteins* **23**, in press.
13 Bonham-Smith, P.C. and Moloney, M.M. (1994) *Plant Physiol.* **106**, 401–402.
14 Li, J., Zhao, J., Rose, A.B., Schmidt, R. and Last, R.L. (1995) *Plant Cell* **7**, 447–461.
15 Hanfstingl, U., Berry, A., Kellogg, E.A., Costa, J.T., Rudiger, W. and Ausubel, F.M. (1994)