HS3D, a Dataset of Homo Sapiens Splice Regions, and its Extraction Procedure from a Major Public Database

Article *in* International Journal of Modern Physics C · December 2002

DOI: 10.1142/S0129183102003796 · Source: CiteSeer

CITATIONS
39

READS
164

2 authors, including:

Salvatore Rampone
Università degli Studi del Sannio
70 PUBLICATIONS **812** CITATIONS

SEE PROFILE

International Journal of Modern Physics C, Vol. 13, No. 8 (2002)

HS^{3}D, A DATA SET OF HOMO SAPIENS SPLICE REGIONS, AND ITS EXTRACTION PROCEDURE FROM A MAJOR PUBLIC DATABASE

PASQUALE POLLASTRO
Facoltà di Scienze, Università del Sannio
Via Port'Arsa 11
Benevento, I-82100, ITALY
E-mail: p.pasquale@tin.it

SALVATORE RAMPONE*
Dpt. Geological and Environmental Studies, Università del Sannio
Via Port'Arsa 11
Benevento, I-82100, ITALY
E-mail: rampone@unisannio.it

Received 1 May 2002
Revised 5 May 2002

The aim of this work is to describe a cleaning procedure of GenBank data, producing material to train and to assess the prediction accuracy of computational approaches for gene characterization. A procedure (GenBank2HS^{3}D) has been defined, producing a dataset (HS^{3}D - Homo Sapiens Splice Sites Dataset) of Homo Sapiens Splice regions extracted from GenBank (Rel.123 at this time). It selects, from the complete GenBank Primate Division, entries of Human Nuclear DNA according with several assessed criteria; then it extracts exons and introns from these entries (actually 4523 + 3802). Donor and acceptor sites are then extracted as windows of 140 nucleotides around each splice site (3799+3799). After discarding windows not including canonical GT–AG junctions (65+74), including insufficient data (not enough material for a 140 nucleotide window) (686+589), including not AGCT bases (29+30), and redundant (218+226), the remaining windows (2796+ 2880) are reported in the data set. Finally, windows of false splice sites are selected by searching canonical GT–AG pairs in not splicing positions (271,937+332,296). The false sites in a range+/- 60 from a true splice site are marked as proximal. HS^{3}D, release 1.2 at this time, is available at the Web server of the University of Sannio: *http://www.sci.unisannio.it/docenti/rampone/*.

Keywords: Extraction algorithm; GenBank; Splice Sites; Data set.

1. Introduction

Genes provide the set of instructions which governs the assembly and function of all human beings. Therefore, gene identification and characterization is a fundamental starting point for describing and understanding our structure, function, and development. However, eukaryotic genomes show a complex underlying structure.

*To whom correspondence should be addressed.

In such genomes a protein-coding gene, started by a promoter, and ended by a poly-A region, consists of a set of regions called exons usually interrupted by other regions called introns. Introns can interrupt both coding (CDS) and noncoding (or untranslated - UTR) regions. The interruption points, Exon – Intron (EI) and Intron - Exon (IE) boundaries, are called donor and acceptor sites, respectively, and in general splice sites. A schematic model of this kind of gene is reported in **Figure 1**.

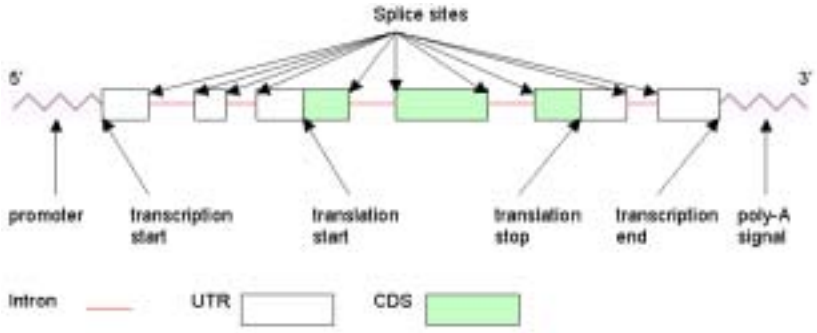

Fig. 1. A schematic gene model.

Introns are removed during a process called splicing. The splicing of introns is part of a multi step process of RNA maturation which takes place in the nucleus to generate mature mRNA molecules for transport to the cytoplasm. This process involves several factors such as snRNP (small nuclear ribonucleoprotein particles) and hnRNPs (heterogeneous nuclear ribonucleoprotein particles). This complex assembly is called spliceosome. The existence is known of consensus sequences around splice sites. For example, the short consensus CAG/G, which is preceded by a region of pyrimidine abundance (polypyrimidine tract), is a typical feature of the acceptor sites.[1] However, a lot of similar sequences are not selected as true splice sites (false positives).

Due to the task complexity, the use of computational methods to predict genes and their structures (sets of spliceable exons) has attracted a considerable research attention in recent years. Many approaches focused on prediction of individual functional elements, e.g. promoters, splice sites, coding regions, in isolation or integrating multiple types of information including splice signal sensors, compositional properties of coding and non-coding DNA and in some cases database homology searching. Some examples of such approaches are NetGene,[2] GeneID,[3] GenMark,[4] FGENEH,[5] Genie,[6] GENESCAN,[7] VEIL,[8] HMMgene,[9] BRAIN,[10] Bayes Networks,[11] GeneFinder,[12] and many others. However the recent release of a draft version of the human genome sequence encoding fewer genes than originally predicted,[13] has led to the suggestion that the computer algorithms used to identify the putative

intron-exon junctions give incorrect estimates more frequently than expected.

As a matter of fact the most current algorithms are based on machine learning approaches.[14] In the machine learning approach, a learning algorithm receives a set of training examples, each labeled as belonging to a particular class (e.g. splicing or not splicing site). The algorithm derives a set parameters from these data and, in some cases, integrates them with a priori knowledge (bias). The algorithm's goal is to produce a classification rule for correctly assigning new examples. The obtained rule is then validated on a control set of data. The success of these methods depends largely on the quality of the data that are used as the training set, since the rule behavior tends to reproduce the training data classification and distribution.[15]

The data are commonly retrieved from three major public databases of annotated DNA sequences: GenBank,[16] EMBL Nucleotide Sequence Database,[17] and DNA Data Bank of Japan (DDBJ). While in other, less extended databases, as ENSEMBL;[a] all available information for a sequence, including proteins, ESTs, mRNAs, promoters, is combined to obtain clean data, the main database sequences, which are mostly compiled from direct author submissions and journal scans, may be found redundant and sometimes inaccurate.

Moreover many authors, in establishing computational methods, select, clean, and process data in very different ways, and a common data set is necessary when the prediction accuracy of different programs needs to be comparatively assessed. Actually, by utilizing new data sets to assess the prediction accuracy of some of these programs, it has been observed that the calculated accuracies were lower than those originally reported by the authors.[18,19]

Finally, the training data must be statistically representative of the problem.[14] For example, to be representative of pentanucleotide composition, even in a uniform distribution hypothesis, the examples in a data set must be of the order of 4^5. Negative examples have also a non trivial role, since the human occurring ratio between positive and negative examples is less than 0.015. There are further needs in the training and control set size ratio.[20]

So there is evidence of several data related problems:

- data may be inaccurate;
- there is cross-correlation between training and control sets due to the repeated sequences present in both the sets;
- the criteria for sequence selection are not always well defined;
- the programs are usually not trained and tested on the same dataset;
- data may be not representative of the problem.

One possible way to avoid these problems is to implement standard datasets and cleaning procedures,[21] but while assessed selection criteria are known and many authors developed their own data, not all are suitable for training and testing advanced machine learning algorithms. For example the Irvine Primate Splice Junc-

[a] http://www.ensembl.org/Homo_sapiens/

tions Dataset (*UCI Machine Learning Repository*[b]) has been for a long time a standard "de facto" in the machine learning community,[22−25] even though it contains errors and does not include sufficient material for the most learning algorithm needs. A more recent and EST confirmed data set[26] has the same limitation in the data extend. Similarly, the limited extend of the HMR195 dataset[c], conceived as a non-training-overlapping control set for gene finding programs,[27] prevents its use as a training set. Recently Burset et al.[28] have developed a much more extensive data base. However, the data do not include false splice sites (negative examples), and, specifically, proximal false splice sites.

The latter data form a well known critical point of classification systems.[19] Actually, while we have no reasons to think that proximal false splice sites are different from distant ones at consensus signal level, we have several evident reasons to think that proximal false splice sites are quite different from distant ones when algorithms recognize them by context information, as triplet or octanucleotide preferences in - also proximal/distant - coding and noncoding regions.[5]

In this paper we describe a new dataset (HS^{3}D - Homo Sapiens Splice Site Dataset) of Homo Sapiens Splice regions extracted from GenBank and its extraction procedure. The aim of this data set is to give standardized material to train and to assess the prediction accuracy of computational approaches for gene characterization.

First the procedure selects, from the GenBank Primate Sequences, entries of Human Nuclear DNA with a complete coding region for at most one gene and with more than one exon, avoiding synthetic, artificial, or foreign genes and pseudogenes or any alternative gene products, conflicts, variations, or mutations in the nucleotide sequence; then it extracts exons and introns from these entries and performs some analyses. Finally it extracts donor and acceptor sites as windows of 140 nucleotides around each splice site. All the windows not including canonical GT–AG junctions, including insufficient data (not enough material for a 140 nucleotide window), including unspecified or erroneous bases, and redundant are discarded.

Finally, windows of false splice sites are selected by searching canonical GT–AG pairs in not splicing positions. The false sites in a range+/- 60 from a true splice site are marked as proximal.

2. Systems and methods

DNA sequence entries are retrieved from GenBank. The GenBank database and related resources are freely accessible via the NCBI home page.[d] There are approximately 13,543,000,000 bases in 12,814,000 sequence records as of August 2001. A new release is made every two months.

The files in the GenBank distribution are divided into 'divisions' that roughly

[b] http://www.ics.uci.edu/∼mlearn/MLRepository.html
[c] http://www.soe.ucsc.edu/∼rogic/evaluation/dataset.html
[d] http://www.ncbi.nlm.nih.gov, and ftp://ncbi.nlm.nih.gov/genbank.

correspond to taxonomic divisions, e.g., Bacteria, Viruses, Primates and Rodents. There are currently 16 divisions. The larger divisions, e.g., EST and Primate, are divided into multiple files. The material treated here comes from the Primate division (files gbpri1.tgz, gbpri2.tgz, ..., gbpriN.tgz). In the Rel.123 it includes 162,557 loci, corresponding to 1,512,130,995 bases.

To develop the HS^{3}D dataset, this material has been algorithmically cleaned and processed as described in the following section.

3. Algorithm

3.1. *Entries Selection*

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references, and a table of features[e] that identifies coding regions and other sites of biological significance, such as transcription units, repeat regions, sites of mutations or modifications and other sequence features. A parsing procedure has been defined using such information to select the entries satisfying the following assessed criteria[21]:

C_I The entries report Human nuclear DNA and not any synthetic, artificial, or foreign genes;

C_{II} The entries do not contain pseudogenes or any alternative gene products, conflicts, variations, or mutations in the nucleotide sequence;

C_{III} The entries contain a complete coding region (complete CDS) for at most one gene and have at least one intron.[f]

3.2. *Exons and Introns*

From each GenBank filtered entry, the introns and exons are extracted and tested. The aim of this step is just to collect exon/intron information, and so no one is discarded. The used rules are the followings:

C_{IV} The original exon/intron number is maintained, if present, or assigned, if this can be made in an unambiguous mode;

C_V The start and the end positions in the original sequence are annotated;

C_{VI} The exon/intron extraction is checked for abnormal termination;

C_{VII} The number of nucleotides for each exon/intron is annotated;

C_{VIII} As several compositional measures depend on the G+C content, such measure is valued for each extracted exon and intron;

C_{IX} Introns usually begin with GT and end with AG dinucleotides. This condition is checked, but no action is taken at this time;

C_X Finally, the presence of nucleotides different from A or G or C or T is tested (nucleotide scan) and annotated;

[e] http://www.ncbi.nlm.nih.gov/collab/FT/index.html

[f] It is worth noting that while we select entries including "Complete CDS" annotation, we do not use their CDS part only, and this implies we potentially select exons of all the 12 Zhang classes[29].

Exons and Introns are reported in the HS^{3}D files *exons.seq* and *introns.seq*. Examples of exons.seq and introns.seq data are reported in **Figures 2 and 3**.

```
Locus : AB000381
Exon N.: 2
Start : 28177
End : 28271
GTCCAGGCTCCTGCGTGAAGTGAT
GCTCCTCTTTGCCTTACTCCTAGCCA
TGGAGCTCCCATTGGTGGCAGCCAGT
GCCACCAT GCGCGCTCAGT
End : Found
Overall nucleotides : 95
G+C content : 60%
Nucleotide scan : Verified
#
```

Fig. 2. Example of exons.seq data.

3.3. *Splice sites*

In the next step splice sites are built as windows of 140 nucleotides, by using adjacent exons/introns materials from the files exons.seq and introns.seq. Then a cleaning process is started and the following entries are removed:

C_{XI} *Non canonical.* Non canonical splice sites are characterized by the absence of canonical consensus di-nucleotide sequences, namely GT and AG (with the introns starting with GT and ending with AG). Such junctions are discarded because this condition may be due to annotation errors.

C_{XII} *Insufficient data sequence.* Entries where there is not enough material for a 140 nucleotide window.

C_{XIII} *Nucleotide scan failed.* This is the case when the presence of nucleotides different from A or G or C or T is detected.

C_{XIV} *Duplicate or redundant.* Windows equal to previously inserted ones (redundant) or both equal and of the same gene tract (duplicated).

The resulting data are reported in the files *EI_true.seq* and *IE_true.seq*. Each file row reports the locus name, the progressive id number of the splice site, the intron number, the exon number, and the 140 nucleotide window (see **Figures 4 and 5**).

3.4. *False Splice Sites*

Finally false splice sites are extracted from the filtered GenBank entries by searching for GT and AG dinucleotides in not splicing positions. Sites falling in categories C_{XII} and C_{XIII} are removed. The sequences having the GT or AG dinucleotide in a range of +/- 60 bp from a true splice site are marked as proximals.

The sequences of false splice sites are reported in the files *EI_false.seq*, and *IE_false.seq*. Each line reports the locus name, the progressive id number of the false

```
Locus : AB000381
Intron N.: 2
Start : 28272
End : 28880
GTAAGTATCATTCCCTCTCACTGTCCTGGAGAGGACGAGAATTCCACCT
GGGGTGCTGGGGGTCACTGGGATGATTGGCTGCAACGTGGAGCAAGCCT
CCGTTAGCTGGGGCCTGCATTGTCTGTGTAATCAGGGGTGGGCACTAGG
GCAGTCCAGGAGTAGTCATGAGCAAGGAGAGGGTTAGGATGAAGGAGCA
GCTGACCAGGGACCAAGGGGGAACCTTGATGTGGCCCTTCCCCATCAGC
GCCAGGCAGGAGGGGCTCTGTCCAGGGAAACCCAGGAGGATGGCGGACC
CCTGTGAGTATCCAGTCTTCCTTGGCGAGGTGAGCCAGGTCTGCAGAGC
ATAGCAATCCCGTATGTGACCACCAAGTGGCGCTCTCTGGAGCCTGCGT
TGGAGAGCAGGGAAAGCTCTCCTTGTGCCTGGCCTCCCTCCCAGGAGCT
AGCCTGGGCCAGACTCAGACTGCATAGAGAGCTGAGCTGTGCAGGCTAG
GAGAAGTCCTTGGAAGCAGAGGGGAAGGGCTGGCCGCTGAAGAAGGGTG
GAGTGAGCTGGTAATGGGTGGAAAAGGCGTAGTGGAGCAGAAGCCTGAAG
CCTGCTTTCTCCCCTCTCAG
End : Found
GT: OK
AG: OK
Overall nucleotides : 609
G+C content : 59.2775041050903%
Nucleotide scan : Verified
#
```

Fig. 3. Example of introns.seq data.

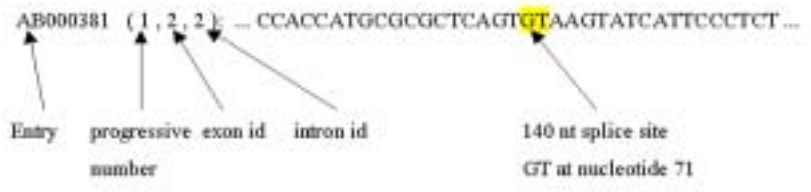

Fig. 4. EI_true.seq entry description.

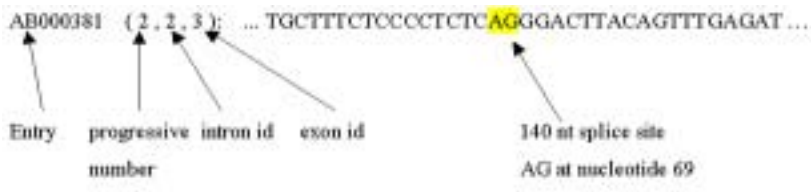

Fig. 5. IE_true.seq entry description.

splice site, the base count of the sequence, the proximal flag (where 1: proximal splice site, and 0: otherwise) and the 140 nt splice site (**Figures 6 and 7**).

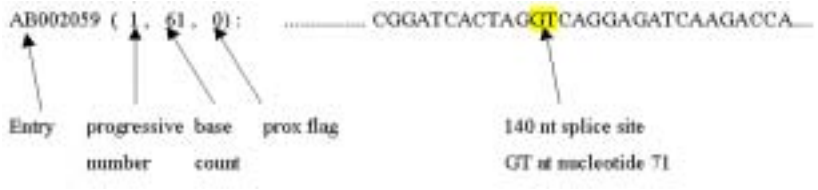

Fig. 6. EI_false.seq entry description.

Fig. 7. IE_false.seq entry description.

The resulting applied procedure is sketched in **Figure 8**.

4. Results

4.1. *Exons and Introns*

The procedure has been applied on the GenBank Release 123. From the complete Primate Sequences (10 gbpriN.tgz files, 162,557 entries), 4523 exons and 3802 introns have been extracted.

Data about exon and intron extraction are reported in the files *exons.stat* and *introns.stat*. Namely the files report the number of extracted exons/introns, the overall nucleotides, the G+C average content, the number of exons/introns that failed the nucleotide scan check, the number of exons/introns in which the annotated end is not found, the minimum, maximum, average, and standard deviation of the exon/intron length. These results are also reported in **Tables 1 and 2.** Exon data must be interpreted as cumulative. In fact the human exons with flanking genomic DNA sequences can be classified into 12 mutually exclusive categories,[29] according to what transcriptional or translational boundaries an exon contains, and each class has its own statistical properties. Namely: (1) a 5Uexon is the 5'-terminal untranslated exon in a gene; (2) a 3Uexon is the 3'-terminal untranslated exon;

Procedure **GenBank2HS^{3}D**
Input: GenBank files gbpri1.tgz, gbpri2.tgz, ..., gbpriN.tgz

1. Select the file entries satisfying C_I, C_{II}, and C_{III};
2. *Exon and Introns*
   - 2.1 Extract the exons and introns from the entries selected in Step 1;
   - 2.2 Perform tests C_{IV}, ..., C_X;
   - 2.3 Build files exons.seq and introns.seq
3. *Splice Sites*
   - 3.1 Build splice sites from adjacent exon/intron couples;
   - 3.2 Remove the sites falling in categories C_{XI}, ..., C_{XIV};
   - 3.3 Build the files EI_true.seq and IE_true.seq;
4. *False Splice Sites*
   - 4.1 Build false splice sites by searching for GT and AG in not splicing positions;
   - 4.2 Remove the sites falling in categories C_{XII}, C_{XIII};
   - 4.3 Mark proximal sites;
   - 4.4 Build the files EI_false.seq and IE_false.seq;

Output: HS^{3}D = {EI_true.seq, IE_true.seq, EI_false.seq and IE_false.seq}.

Fig. 8. GenBank2HS^{3}D procedure schema.

Table 1. Exon statistics

Extracted Exons	4523
Overall nucleotides	1062158
G+C average content	55.31%
Nucleotide scan failed	16
End not found	0
Min length	5
Max length	9310
Average length	234. 83
Standard deviation	429.77

Table 2. Introns statistics

Extracted Introns	3802
Overall nucleotides	4411560
G+C average content	52.34%
Nucleotide scan failed	195
End not found	0
Not canonical start	74
Not canonical end	81
Min length	10
Max length	91801
Average length	1160.32
Standard deviation	3506.86

(3) a 5UTexon is the 5'-terminal exon having a 5'UTR followed by a CDS; (4) a 3TUexon is the 3'-terminal exon having a 3'UTR following a CDS; (5) an IUTexon is an internal exon having a 3' portion of the 5'UTR followed by a CDS; (6) an ITUexon is an internal exon having a 5' portion of 3'UTR following a CDS; (7) an IUexon is an internal untranslated exon; (8) an ITexon is an internal translated exon; (9) a 5UTUexon does not contain the transcriptional end; (10) a 3UTUexon does not contain the transcriptional start; (11) a 5-3UTUexon contains both; and (12) an IUTUexon contains neither.

To verify the congruence of the obtained data to known properties,[29] we divide exons into three main categories: (1) *Initial*, corresponding to the first exon on the 5' side, (2) *Terminal*, the 3'ending one, and (3) *Internal*, covering all the remaining ones. **Figure 9** shows the HS^{3}D exon length distribution for each category. The exon length is indicated on the x axis (logarithmic scale), and the number of exons on the y one. As expected, Internal exons, where ITexons are the majority, have a log-normal distribution centered around 130 nt, while initial exons, mostly 5Uexons and 5UTexons, are relatively short ($\leq$ 100 nt), and terminal exons, mostly 3TUexons and 3Uexons, relatively long ($\geq$ 100 nt) and extremely heterogeneous.

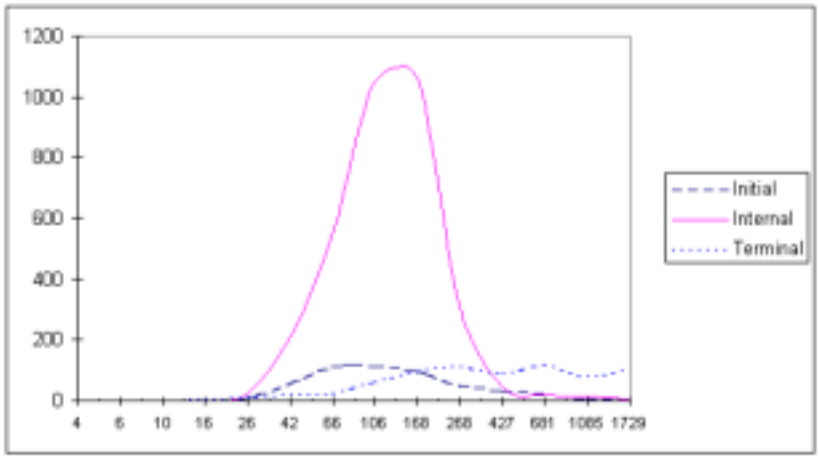

Fig. 9. Exon Size Distribution.

4.2. *Splice Sites*

In the procedure next step 3799 + 3799 donor and acceptor sites have been extracted as windows of 140 nucleotides around each splice site. After discarding sequences not including canonical GT–AG junctions (65+74), including insufficient data (not enough material for a 140 nucleotide window) (686+589), including not AGCT bases (29+30), and redundant (218+226), there are 2796+ 2880 windows.

The files *EI_true.inf,* and *IE_true.inf* report, for each locus, details about the extracted splice sites.

The files *EI_true.stat*, and *IE_true.stat* report some statistics, namely: the number of splice sites extracted, the number of non canonical splice sites, the number of splice sites with insufficient data sequence, the number of splice sites that failed the nucleotide scan check, the number of redundancies, and the overall number of splice sites. Such statistics are also reported in **Tables 3** and **4**.

Table 3. Exon-Intron true splice sites statistics

Extracted EI splice sites	2796
Not canonical found	65
Insufficient data sequence	686
Nucleotide scan failed	29
Redundant	218
Total	3799

Table 4. Intron-Exon true splice site statistics

Extracted IE splice sites	2880
Not canonical found	74
Insufficient data sequence	589
Nucleotide scan failed	30
Redundant	226
Total	3799

4.3. *False Splice Sites*

In the last step 287,296+348,370 windows of false splice sites are selected. The false sites in a range+/- 60 from a true splice site are marked as proximal. Data about false splice sites extracted and discarded, the latter mostly by insufficient data sequence, are reported in the files *EI_false.inf*, and *IE_false.inf*.

4.4. *Server Data*

The resulting HS^{3}D dataset is available at the Web server of the University of Sannio[9] The described data are downloadable in compressed format (.zip). To simplify the downloading, false splice sites data are divided into 3 + 4 files, and a merging utility is provided. **Table 5** reports the data organization.

5. Discussion

The aim of this work was to define a cleaning procedure of GenBank data, producing standardized material to train and to assess the prediction accuracy of computational approaches for gene characterization.

The defined procedure (GenBank2HS^{3}D), based on 14 constraints and selection criteria, allows the extraction of a dataset from GenBank in a one shot, fully au-

[9]http://www.sci.unisannio.it/docenti/rampone/

Table 5. Server Data Organization

File	Content	Description
exons.zip	exons.seq.	Exon sequences.
introns.zip	introns.seq.	Intron sequences.
EI_true.zip	EI_true.seq, EI_true.inf.	Exon-Intron true splice sites and informations.
IE_true.zip	IE_true.seq, IE_true.inf.	Intron-Exon true splice sites and informations.
EI_false_1.zip	EI_false.seq.001, EI_false.inf, EI_false.seq.bat.	Exon-Intron false splice sites (Part 1), informations, and merging utility.
EI_false_2.zip	EI_false.seq.002.	Exon-Intron false splice sites (Part 2).
EI_false_3.zip	EI_false.seq.003.	Exon-Intron false splice sites (Part 3).
IE_false_1.zip	IE_false.seq.001, IE_false.inf, IE_false.seq.bat.	Intron-Exon false splice sites (Part 1), informations, and merging utility.
IE_false_2.zip	IE_false.seq.002.	Intron-Exon false splice sites (Part 2).
IE_false_3.zip	IE_false.seq.003.	Intron-Exon false splice sites (Part 3).
IE_false_4.zip	IE_false.seq.004.	Intron-Exon false splice sites (Part 4).
statistics.zip	exons.stat, introns.stat, EI_true.stat, IE_true.stat.	Statistics.

tomated, easy upgradeable way. The use of detailed intermediate results allows to trace the extraction steps.

The extraction choices are intended to balance two opposite necessities. On the one hand they aim at minimizing the probability of error, on the other hand they aim at preserving a statistically meaningful quantity of data. Unlike other authors, we have chosen to discard sites rather than to correct errors, though trivial, in the sequences, in order to avoid the risk of introducing new ones. For the same reason, non canonical splicing sites have not been included, even if these comprise about the 3% of the GenBank annotated splice pairs. Moreover, we have avoided comparisons with pre-processed data and ESTs databases, that would have involved further losses of data or the necessity of verifications of reliability. Altogether these choices lead us to maintain about 70% of the sites of splicing initially selected.

The resulting HS^{3}D, Homo Sapiens Splice Site Dataset, having data extent of about 650,000 true/false donor and acceptor sites, is able to satisfy the most current algorithms needs.

Actually we do not include sites from alternatively spliced genes. In this case, since alternative splicing in human genome is not uncommon, these data will be introduced in a following data set release. However, alternative cases appear to have separate rules, and they will have a separate treatment (and, obviously, a different ratio between the discarded data and the risk).

Acknowledgements

The authors wish to thank Francesco P. Mancini for useful discussions, and Massimo Mastroianni, Webmaster of the Facoltà di Scienze, Università del Sannio, for his patience.

References

1. Senapathy P., Shapiro M.B., and Harris N.L. (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods in Enzymology*, **183**, 252-278.
2. Brunak S., Engelbrecht J., and Knudsen S. (1991) Prediction of the human mRNA donor and acceptor sites from the DNA Sequence. *J.Mol.Biol.*, **220**, 49-65.
3. Guigo R., Knudsen S., Drake N., and Smith T. (1992) Prediction of gene structure. *J.Mol.Biol.* **226**, 141-157.
4. Borodovsky M., and McIninch J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems,* **30**, 161-171.
5. Solovyev V.V., Salamov A.A., and Lawrence C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, **22**, 5156-5163.
6. Kulp D., Haussler D., Reese M.G., and Eeckman F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int. Conf. Intell. Syst. Mol. Biol.*, 4, pp. 134-142.
7. Burge C., and Karlin S. (1997) Prediction of complete gene structure in human genomic DNA. *J.Mol.Biol.*, **268**, 78-94.
8. Henderson J., Salzberg S., and Fasman K.H. (1997) Finding Genes in DNA with a Hidden Markov Model. *J.Comput.Biol.,* **4**, 127-141.
9. Krogh A. (1998) An Introduction to Hidden Markov Models for Biological Sequences. In Salzberg, S.L., Searls, D.B., and Kasif, S. (eds),*Computational methods in Molecular Biology*, Elsevier, pp. 45-63.
10. Rampone S. (1998) Recognition of Splice-Junctions on DNA Sequences by BRAIN learning algorithm. *Bioinformatics*, **14**, 676-684.
11. Cai D., Delcher A., Kao B., and Kasif S. (2000) Modelling splice sites with Bayes Networks. *Bioinformatics*, **16**, 152-158.
12. Pertea M., Lin X., and Salzberg S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research*, **29**, 1185-1190.
13. Venter J.C et al. (2001) The sequence of the human genome. *Science*, **291**, 1304-1351 (Erratum in: Science Jun 5; **292**, 1838).
14. Baldi P., and Brunak S., (1998) *Bioinformatics: The Machine Learning Approach.* MIT Press.
15. Bishop C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press.
16. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., and Wheeler D.L. (2000) GenBank. *Nucleic Acids Research,* **28**, 15-18.
17. Stoesser G., Baker W., van den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Lombard V., Lopez R., Parkinson H., Redaschi N., Sterk P., Stoehr P., and Tuli M.A. (2001). The EMBL nucleotide sequence database. *Nucleic Acids Research,* **29**, 17-21.
18. Burset M., and Guigo R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353-367.
19. Thanaraj T.A. (2000) Positional Characterisation of False Positives from Computational Prediction of Human Splice Sites. *Nucleic Acids Research*, **28**, 744-754.
20. Niyogi P., and Girosi F. (1996) On the Relationship Between Generalization Error, Hypothesis Complexity, and Sample Complexity for Radial Basis Functions. *Neural Computation*, **8**, pp. 819-842.
21. Thanaraj T.A. (1999a) Standards to Create Clean Data Sets for Gene Prediction. *Bioinformer*, Fall '99.[h]

[h]http://bioinformer.ebi.ac.uk/newsletter/archives/5/gene_prediction.html

22. Noordewier M.O., Towell G.G., and Shavlik J.W. (1991) Training Knowledge-Based Neural Networks to Recognize Genes in DNA Sequences. *Advances in Neural Information Processing Systems*, 3, Morgan Kaufmann.
23. Towell G.G. (1991) Symbolic Knowledge and Neural Networks: Insertion, Refinement, and Extraction. *PhD Thesis*, University of Wisconsin - Madison.
24. Towell G.G., and Shavlik J.W. (1992) Interpretation of Artificial Neural Networks: Mapping Knowledge-based Neural Networks into Rules. In *Advances in Neural Information Processing Systems*, 4, Morgan Kaufmann.
25. Towell G.G., Shavlik J.W., and Craven M.W. (1991) Constructive Induction in Knowledge-Based Neural Networks. In *Proceedings of the Eighth International Machine Learning Workshop*, Morgan Kaufmann.
26. Thanaraj T.A. (1999b) A Clean data set of EST-confirmed Splice Sites from Homo Sapiens and Standards for Clean-up Procedures. *Nucleic Acids Research*, **27**, 2627-2637.
27. Rogic S., Mackworth A.K., and Ouellette F.B.F. (2001) Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Res.* **11**, 817-832.
28. Burset M., Seledtsov I.A., and Solovyev V.V. (2001) SpliceDB: dataset of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research*, **29**, 255-259.
29. Zhang M.Q. (1998) Statistical Features of Human Exons and their Flanking Regions. *Human Molecular Genetics*, **7**, 919-932.