BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database
Volume 8(5)

IntDb: A comprehensive database for classified introns of saccharomyces & human

Subhalaxmi Mohanty, Amouda Nizam* & Monalisha Biswal

Centre for Bioinformatics, School of Life Sciences, Pondicherry University, R.V. Nagar, Kalapet, Puducherry - 605014; Amouda Nizam – Email: amouda@yahoo.com; Phone: +91-413-2655212; Fax: +91-413-2655211; *Corresponding author

Received February 22, 2012; Accepted March 03, 2012; Published March 17, 2012

Abstract:
Introns (intra-genic) are non-coding regions of several eukaryotic genes. However, their role in regulation of transcription, embryonic development, stimulate gene (HEG) is apparent in recent years. Thus current research focuses on mutation in introns and their influence in causing various diseases. Though many available intron databases like YIDB, IDB, ExInt, GISSD, FUGOID, etc. discusses on various aspects of introns but none of them have classified the introns where identification of start intron is found to be important which mainly regulates the various activities of protein at gene level. This lead to an idea for development of "Intdb"; a database meant for classifying introns as start, middle and stop on the basis of position of specific consensus site. Information provided in IntDb is useful for gene prediction, determination of splicing sites and identification of diseases. In addition, the main focus is on violation of consensus rule and frequency of other deviations observed in classified introns. Further, GC content, length variations according to the biased residues and occurrence of consensus pattern to discover potential role of introns is also emphasized in IntDb.

Availability: http://introndb.bicpu.edu.in/

Keywords: consensus pattern, consensus rule violation, length variation, splicing sites.

Background:
Introns (intra-genic, non-coding DNA) split a eukaryotic gene into coding exons and their role largely remains unknown. Variation of intron length with respect to intron positions and its dynamic study is focused by current researchers **[1].** A correlation between intron positions and protein structure is the main theme of research after the discovery of intron in eukaryotes in 1977 **[2].** It is found that there was a strong positive correlation between intron length and divergence, also a strong negative correlation between intron length and GC content (first introns are more GC rich, longer, more divergent) **[3].** A major type of alternative splicing mechanism in human transcriptome is intron retention. In 21,106 human genes intron retention is located within the untranslated regions (UTRs) of human transcript and 22% of retained introns in human are also found in mouse transcriptome **[4].** In upstream and downstream introns, flanking intronic sequences are 88% and 80% respectively which is higher in comparison to promoter region conservation level as 77% **[5].** According to the mechanism of excision introns are classified as GroupI, GroupII, GroupIII, t-RNA, spliceosomal, chloroplast, mitochondrial, pre-mRNA and HAC1 introns **[6].** We have developed IntDb, a comprehensive database which contains information regarding accession ID, definition, location, length, nucleotide sequences of classified introns (like start, middle, stop) of *Saccharomyces* and human. IntDb constitutes percentage of GC content, GT-AG, AT-AC consensus, etc. along with other deviations and information related to diseases like application of homing endonucleases in human. IntDb has better constitution than other intron databases like YIDB, FUGOID, ExInt, GISSD, IDB

in exploring classified introns and their analysis **Table 1 (see supplementary material).**

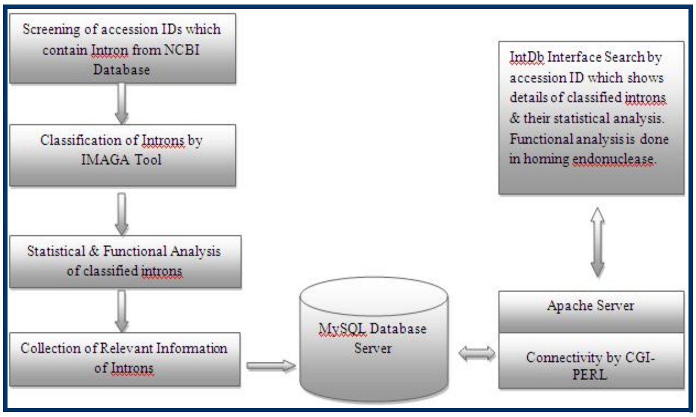
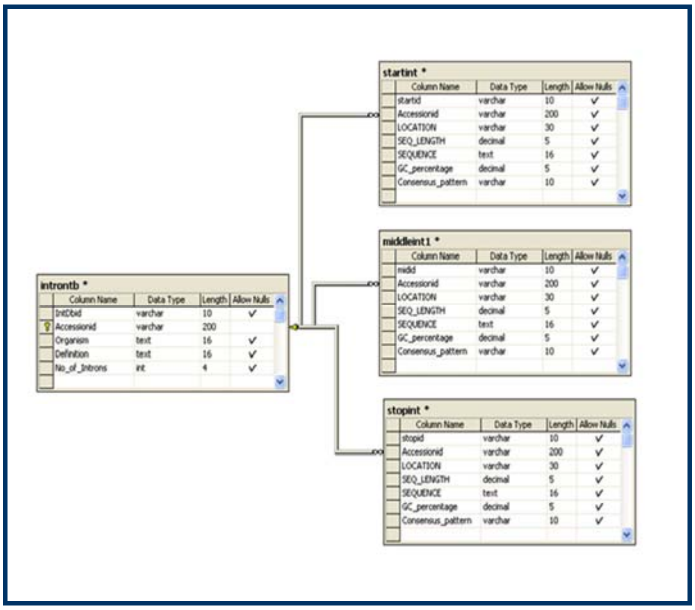

Screening of accession IDs which contain Intron from NCBI Database
Classification of Introns by IMAGA Tool
Statistical & Functional Analysis of classified introns
Collection of Relevant Information of Introns
MySQL Database Server
Connectivity by CGI-PERL
Apache Server
IntDb Interface Search by accession ID which shows details of classified introns & their statistical analysis. Functional analysis is done in homing endonuclease.

Figure 1: Flow chart of IntDb Development

Methodology:

IntDb is developed by using HTML **(Figure 1)**, Java script as frontend, MySQL as backend and CGI-PERL for connectivity to retrieve data. Apache server which has most intuitive configuration is used as the web-server. The complete genome of human & Saccharomycetaceae family is screened in NCBI database to find accession IDs (Genbank sequences) those posses introns and after screening 18 species of Saccharomycetaceae family, only 3 species like *Saccharomyces cerevisiae* (350 IDs), *S.pastorias* (7 IDs), *S.bayanus* (8 IDs) were found to be containing introns. Along with this 500 *Homo sapiens* accession IDs those posses' introns are collected. Classified introns are collected by using IMAGA tool which classifies the intron as start, middle and stop by taking introns positions from Genbank file. The database constitution is presented in the schema **(Figure 2)**. Patterns of introns were found by using IMAGA tool (http://imaga.bicpu.edu.in:8080/Isaga/) in which 4 sets of introns taken in one slots with variation of GA arameters (population size, maximum number of GA runs, crossover type).



startint *

Column Name	Data Type	Length	Allow Nulls
startid	varchar	10	✓
Accessionid	varchar	200	✓
LOCATION	varchar	30	✓
SEQ_LENGTH	decimal	5	✓
SEQUENCE	text	16	✓
GC_percentage	decimal	5	✓
Consensus_pattern	varchar	10	✓

middleint1 *

Column Name	Data Type	Length	Allow Nulls
midid	varchar	10	✓
Accessionid	varchar	200	✓
LOCATION	varchar	30	✓
SEQ_LENGTH	decimal	5	✓
SEQUENCE	text	16	✓
GC_percentage	decimal	5	✓
Consensus_pattern	varchar	10	✓

introntb *

Column Name	Data Type	Length	Allow Nulls
IntDbid	varchar	10	✓
Accessionid	varchar	200	
Organism	text	16	✓
Definition	text	16	✓
No_of_Introns	int	4	✓

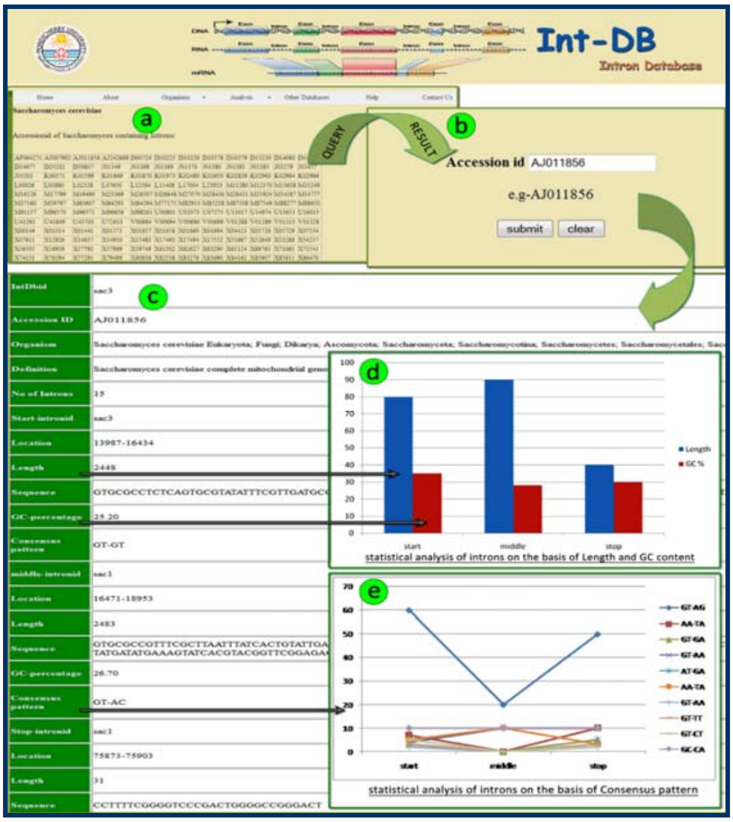
stopint *

Column Name	Data Type	Length	Allow Nulls
stopid	varchar	10	✓
Accessionid	varchar	200	✓
LOCATION	varchar	30	✓
SEQ_LENGTH	decimal	5	✓
SEQUENCE	text	16	✓
GC_percentage	decimal	5	✓
Consensus_pattern	varchar	10	✓

Figure 2: Schema of IntDb shows 4 tables

Utility:

IntDb constitutes classified introns of *Saccharomyces* and human along with accession IDs those posses' introns with relevant information regarding introns like definition, no. of introns, location, length, nucleotide sequence, GC percentage and consensus pattern of classified introns **(Figure 3a, 3b, 3c).** We have analyzed several statistical and functional aspects of introns. Statistical analysis is done by considering 850 accession IDs of Sachharomyces (350) and human (500). Statistical analysis is done to identify consensus pattern, GC content and length variation in classified introns **(Figure 3d, 3e).** In these introns containing accession IDs first, last introns are said as start, stop intron and rest all in between are middle introns. The collected accession IDs always posses at least one start intron, hence start introns are more. This analysis is done by taking 850 start, 38 middle and 143 stop introns. Apart from this we have found some patterns of start introns by using IMAGA tool. All the previous research works show that first introns (start intron) are longer but our analysis reflects middle introns are longer in average though their occurence is less in gene. Functional analysis of introns covers the detailed prior studies on introns, their role in homing endonuclease and transcription. Here homing endonuclease application is projected in diseases like Xeroderma pigmentosa and cancer. Transcription regulatory role is found in Murine IL4, Hsp90 beta gene, Beta 1 tubulin gene in which intron enhances expression of gene by Intron Mediated enhancement (IME). IntDb can be accessed for detection of mutation position, identification of splicing sites which are helpful in gene prediction programs, development of tools & databases.



Accession id AJ011856

e.g-AJ011856

IntDbid	sac3
Accession ID	AJ011856
Organism	Saccharomyces cerevisiae Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomyceta; Saccharomycotina; Saccharomycetes; Saccharomycetales; Sac
Definition	Saccharomyces cerevisiae complete mitochondrial geno
No of Introns	15
Start-intronid	sac3
Location	13987-16434
Length	2448
Sequence	GTGCGCCTCTCAGTGCGTATATTTCGTTGATGC
GC-percentage	25.20
Consensus pattern	GT-GT
middle-intronid	sac1
Location	16471-18953
Length	2483
Sequence	GTGCGCCGTTTCGCTTAATTTATCACTGTATTGA TATGATATGAAAGTATCACGTACGGTTCGGAGA
GC-percentage	26.70
Consensus pattern	GT-AC
Stop-intronid	sac1
Location	75873-75903
Length	31
Sequence	CCTTTTCGGGGTCCCGACTGGGGCCGGGACT

Figure 3: A snapshot of IntDb and its important options. **(a)** Table contains accessionid of Saccharomyces cerevisiae possess introns, **(b)** Search for accession ID AJ011856 in IntDb homepage, **(c)** Result for query AJ011856, **(d)** Statistical analysis

of start, middle, stop introns for its GC content and length, and **(e)** Statistical analysis for consensus pattern of start, middle, stop introns.

Future Developments:

IndDb is a generalized name and intron data of other model organisms will be included in future. We also want to predict and store typical patterns which are conserved during evolution along with their functional or malfunctional role in gene.

Acknowledgement:

The authors would like to thank the Department of Biotechnology (DBT), Government of India for providing fellowship and financial support.

References:

[1] Gazave E *et al*. *Genome Biol*. 2007 **8**: R21 [PMID: 17309804]

[2] Whamond GS & Thornton JM, *et al. J Mol Biol*. 2006 **359**: 238 [PMID: 16616935]

[3] Bradnam KR & Korf I *et al*. *PLoS One*. 2008 **3**: e3093 [PMID: 18769727]

[4] Galante PA *et al*. *RNA*. 2004 **10**: 757 [PMID: 15100430]

[5] Sorek R & Ast G, *et al. Genome Res*. 2003 **13**: 1631 [PMID: 12840041]

[6] Lopez PJ *et al. Nucleic Acids Res*. 2000 **28**: 85 [PMID: 10592188]

Edited by P Kangueane

Citation: **Mohanty** *et al*. Bioinformation 8(5): 233-236 (2012)

Supplementary Material:

Table 1: Comparison of IntDb with other Intron databases

Database	About the database	URL	Classified Introns	Consensus pattern	Statistical Analysis	Functional Analysis
IntDb	It includes 850 classified Introns of Saccharomyces and Human	http://introndb.bicpu.edu.in/	It constitutes classified introns (start, middle, stop)	Different types of pattern like GT-AG, AA-TA, GT-GA, GT-AA, AT-GA, AA-TA, GT-AA, GT-TT, GT-CT, GC-CA are shown here.	Length and GC % of classified introns and ratio of start, middle, stop intron are compared among them	Roles of intron in transcription, homing endonuclease and maturase activity.
YIDB	254 nuclear and mitochondrial intron	http://compbio.soe.ucsc.edu/yeast_introns.html	Not available	Not available	Splice sites and intron length are provided but not GC %	Not available
FUGOID	354 introns	http://fugoid.webhost.utexas.edu/introndata/main.htm	Not available	Not available	Length are given but not GC %	Not available
ExInt	5, 25, 870 introns	http://intron.bic.nus.edu.sg/exint/exint.html	Not available	Not available	Intron length and phase distribution are given but not GC%	Not available
GISSD	1, 789 Gr-I introns	http://www.rna.whu.edu.cn/gissd/	Not available	Not available	Only classifies Gr-I intron into 14 subgroups	Not available
IDB	1, 54, 000 introns	http://nutmeg.bio.indiana.edu/intron/index.html	Not available	Not available	Length and GC % of introns not classified introns are given	Not available