

A New File Format and Tools for the Large-Scale Data Submission to DNA Data Bank of Japan (DDBJ)

Satoru Miyazaki¹ Hiroyuki Hashimoto² Akemi Shimada¹
smiyazak@genes.nig.ac.jp hirohash@genes.nig.ac.jp akshimad@genes.nig.ac.jp
Yoshio Tateno¹ Hideaki Sugawara¹
ytateno@genes.nig.ac.jp hsugawar@genes.nig.ac.jp

¹ Center for Information Biology, National Institute of Genetics,
1111 Yata, Mishima, Shizuoka 411-8540, Japan

² Hitachi Software Engineering Co., Ltd. Yokohama 231-0015, Japan

Keywords: DNA sequence, genome, large-scale submission, JAVA, object-oriented

1 Introduction

Thanks to the genome projects and the rapid development of high-throughput technique to determine DNA sequences, two types of very large-scale data have been submitted to DDBJ/EMBL/Genbank International Nucleotide Sequence Databases (INSD). One is a whole genome sequence of an organism. Another is a set of sequences for the phylogenetic study based on a ubiquitous gene among hundreds of organisms. It is very important for researchers to get accession numbers of their sequences in a timely fashion. INSD has to meet their needs even if the data is massive.

To make submission of data easier, INSD has developed and provided such tools for stand-alone usage and Web submission as Authorin, Sequin, SAKURA, Webin, BankIt. These submission tools, however, target relatively small scale data and are not suitable for massive data.

Therefore we at DDBJ developed a new data file format and off-line tools to support the submission of the large-scale sequence data.

2 Method and Results

2.1 A new file format

Authorin and Sequin use the transaction file format and ASN.1 format respectively. EMBL flat file format is required to send your data to EMBL database. These formats, however, have some points we have to consider. For instance, files described by ASN.1 format are difficult to edit directly, but easy to find some syntax error by computers. EMBL format is easy to edit by yourself, but difficult to describe many kinds of combination of feature keys and qualifiers, that are two level data descriptors, without syntax errors. To realize readability and easiness to parse, we designed a new file format for the description of biological annotation to sequences. It is a tab-delimited text format with five columns (Entry, Feature, Location, Qualifier and Value) (See Fig. 1).

2.2 Mass Submission Tool (MST)

MST is a stand-alone application program and is able to handle the annotation data consisting of over 100 features with various qualifiers for more than 10,000 sequences. This application consists of three modules: annotation editor, sequence editor and parser tool. MST is developed based on the object-oriented concept by use of a library implemented by JAVA language. The features from the viewpoints of programming are:

1. executable on all-most all operating system
2. easy to modify user interface for the addition and the deletion of feature keys and qualifiers
3. each module is also executable as an application by minor modification of source code

Fig. 2 is sample windows of annotation editor on MST.

Line	Feature Key	Qualifier	Value
1	CDS	join(1)	1..1000
2	CDS	join(2)	1001..2000
3	CDS	join(3)	2001..3000
4	CDS	join(4)	3001..4000
5	CDS	join(5)	4001..5000
6	CDS	join(6)	5001..6000
7	CDS	join(7)	6001..7000
8	CDS	join(8)	7001..8000
9	CDS	join(9)	8001..9000
10	CDS	join(10)	9001..10000
11	CDS	join(11)	10001..11000
12	CDS	join(12)	11001..12000
13	CDS	join(13)	12001..13000
14	CDS	join(14)	13001..14000
15	CDS	join(15)	14001..15000
16	CDS	join(16)	15001..16000
17	CDS	join(17)	16001..17000
18	CDS	join(18)	17001..18000
19	CDS	join(19)	18001..19000
20	CDS	join(20)	19001..20000
21	CDS	join(21)	20001..21000
22	CDS	join(22)	21001..22000
23	CDS	join(23)	22001..23000
24	CDS	join(24)	23001..24000
25	CDS	join(25)	24001..25000
26	CDS	join(26)	25001..26000
27	CDS	join(27)	26001..27000
28	CDS	join(28)	27001..28000
29	CDS	join(29)	28001..29000
30	CDS	join(30)	29001..30000
31	CDS	join(31)	30001..31000
32	CDS	join(32)	31001..32000
33	CDS	join(33)	32001..33000
34	CDS	join(34)	33001..34000
35	CDS	join(35)	34001..35000
36	CDS	join(36)	35001..36000
37	CDS	join(37)	36001..37000
38	CDS	join(38)	37001..38000
39	CDS	join(39)	38001..39000
40	CDS	join(40)	39001..40000
41	CDS	join(41)	40001..41000
42	CDS	join(42)	41001..42000
43	CDS	join(43)	42001..43000
44	CDS	join(44)	43001..44000
45	CDS	join(45)	44001..45000
46	CDS	join(46)	45001..46000
47	CDS	join(47)	46001..47000
48	CDS	join(48)	47001..48000
49	CDS	join(49)	48001..49000
50	CDS	join(50)	49001..50000
51	CDS	join(51)	50001..51000
52	CDS	join(52)	51001..52000
53	CDS	join(53)	52001..53000
54	CDS	join(54)	53001..54000
55	CDS	join(55)	54001..55000
56	CDS	join(56)	55001..56000
57	CDS	join(57)	56001..57000
58	CDS	join(58)	57001..58000
59	CDS	join(59)	58001..59000
60	CDS	join(60)	59001..60000
61	CDS	join(61)	60001..61000
62	CDS	join(62)	61001..62000
63	CDS	join(63)	62001..63000
64	CDS	join(64)	63001..64000
65	CDS	join(65)	64001..65000
66	CDS	join(66)	65001..66000
67	CDS	join(67)	66001..67000
68	CDS	join(68)	67001..68000
69	CDS	join(69)	68001..69000
70	CDS	join(70)	69001..70000
71	CDS	join(71)	70001..71000
72	CDS	join(72)	71001..72000
73	CDS	join(73)	72001..73000
74	CDS	join(74)	73001..74000
75	CDS	join(75)	74001..75000
76	CDS	join(76)	75001..76000
77	CDS	join(77)	76001..77000
78	CDS	join(78)	77001..78000
79	CDS	join(79)	78001..79000
80	CDS	join(80)	79001..80000
81	CDS	join(81)	80001..81000
82	CDS	join(82)	81001..82000
83	CDS	join(83)	82001..83000
84	CDS	join(84)	83001..84000
85	CDS	join(85)	84001..85000
86	CDS	join(86)	85001..86000
87	CDS	join(87)	86001..87000
88	CDS	join(88)	87001..88000
89	CDS	join(89)	88001..89000
90	CDS	join(90)	89001..90000
91	CDS	join(91)	90001..91000
92	CDS	join(92)	91001..92000
93	CDS	join(93)	92001..93000
94	CDS	join(94)	93001..94000
95	CDS	join(95)	94001..95000
96	CDS	join(96)	95001..96000
97	CDS	join(97)	96001..97000
98	CDS	join(98)	97001..98000
99	CDS	join(99)	98001..99000
100	CDS	join(100)	99001..100000

Figure 1: The format of Annotation file

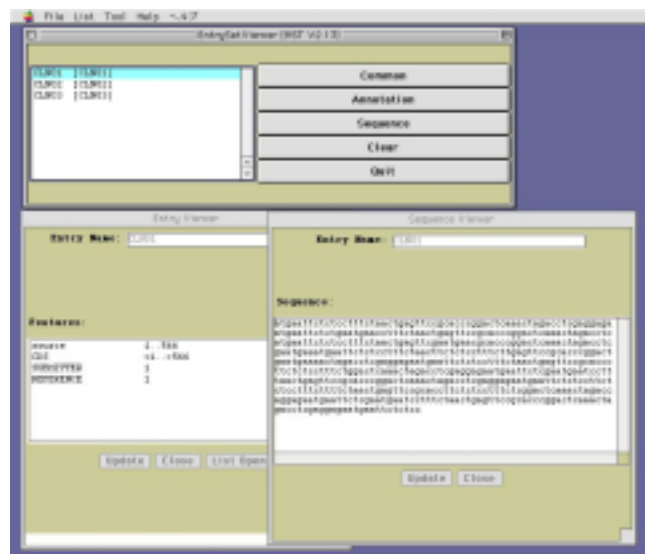


Figure 2: Sample windows on MST

References

- [1] Sugawara, H., Miyazaki, S., Gojobori, T. and Tateno, Y., DNA Data Bank of Japan dealing with large-scale data submission, *Nucleic Acids Res.*, 27(1):25–28, 1999.
- [2] Tateno, Y., Fukami-Kobayashi, K., Miyazaki, S., Sugawara, H. and Gojobori, T., DNA Data Bank of Japan at work on genome sequence data, *Nucleic Acids Res.*, 26(1):16–20, 1998.
- [3] Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. and Gojobori, T., DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams [In Process Citation], *Nucleic Acids Res.*, 28(1):24–26, 2000.