

1           **GenBank 2 Sequin - a file converter preparing custom GenBank files for database submission**

2   Pascal Lehwark<sup>1</sup>, Stephan Greiner<sup>2\*</sup>

3  
4     <sup>1</sup> Die Freiraum.Company Web & Code UG, Glogauer Straße 31, 10999 Berlin, Germany,

5     <sup>2</sup> Max Planck Institute of Molecular Plant Physiology, Department for Organelle Biology, Biotechnology  
6     and Molecular Ecophysiology, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

7  
8     E-mail:

9     [pascal@ntzwrk.de](mailto:pascal@ntzwrk.de) (PL)

10    [greiner@mpimp-golm.mpg.de](mailto:greiner@mpimp-golm.mpg.de) (SG)

11  
12    \* Corresponding author

13 **Abstract**

14 The typical wet lab user often annotates smaller sequences in the GenBank format, but resulting files  
15 are not accepted for database submission by NCBI. This makes submission of such annotations a  
16 cumbersome task. Here we present “GenBank 2 Sequin” an easy-to-use web application that converts  
17 custom annotations in the GenBank format into the NCBI direct submission format Sequin. Additionally,  
18 the program generates a “five-column, tab-delimited feature table” and a FASTA file required for  
19 submission through BankIt or for the update an existing GenBank entry. We specifically developed  
20 “GenBank 2 Sequin” for the regular wet lab researcher with strong focus on user-friendliness and  
21 flexibility. It is equipped with an intuitive graphical interface and a comprehensive documentation. The  
22 application can be employed to prepare any GenBank file for database submission. It is freely available  
23 online at <https://chlorobox.mpimp-golm.mpg.de/GenBank2Sequin.html>.

24

25 **Keywords:** NCBI, GenBank, Database Submission, Web Server

## 26 **Background**

27 The typical wet lab user often annotates smaller sequences such plasmids with commercial sequence  
28 visualization and annotation software like Vector NTI Advance (Life Technologies, Invitrogen, Carlsbad,  
29 CA, USA) or Lasergene SeqBuilder (DNASTAR, Madison, WI, USA). The resulting GenBank or EMBL files,  
30 however, are not accepted for submission by NCBI. NCBI itself provides the web-based tool BankIt or the  
31 stand-alone programs Sequin and tbl2asn as annotation and/or submission tools [1], but again, these  
32 programs also do not read GenBank or EMBL files. Instead, the user must provide a so-called “five-  
33 column, tab-delimited feature table” (<http://www.ncbi.nlm.nih.gov/Sequin/table.html>) in order to avoid  
34 time-consuming manual feature input. However, one needs to substantially familiarize with the NCBI  
35 syntax to create such an annotation table from a GenBank entry. Moreover, tbl2asn, the powerful  
36 command line program of NCBI that creates Sequin files suitable for submission, requires both, an  
37 annotation table and some computational skills.

38 Unfortunately, the only public browser based file converters which generate Sequin files or  
39 annotation tables from GenBank entries (gbk2sqn and gbk2tbl; developed by Andre Villegas and Paulina  
40 Konczy, Laboratory for Foodborne Zoonoses, Guelph, ON, Canada) are no longer supported [2]. A  
41 GenBank parser (gbf2tbl.pl) provided by NCBI  
42 ([ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi\\_tools/converters/scripts/gbf2tbl.pl](ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/scripts/gbf2tbl.pl)) can only partially replace  
43 the two programs. Similar to our tool described below, the script produces annotation tables and FASTA  
44 files from GenBank records. These files can subsequently be read by tbl2asn to create Sequin files for  
45 direct submission. The GenBank parser, however, is not user friendly. It is only provided as a perl script  
46 and tbl2asn must be manually executed. Last, implemented features in free standalone programs such  
47 as Artemis [3], that convert GenBank files into submission formats, require installation of these  
48 additional software.

49 In summary, there is a strong demand for a web-based, easy-to-use file converter, which directly  
50 converts GenBank annotations into “five-column, tab-delimited feature tables” and additionally  
51 provides Sequin files for direct submission. Therefore, we developed GenBank 2 Sequin, as part of the  
52 CHLOROBX toolkit (<https://chlorobox.mpimp-golm.mpg.de>), hosted and developed at the Max Planck  
53 Institute of Molecular Plant Physiology (Potsdam/Golm, Germany). The toolbox offers software  
54 applications for the analysis of (plant derived) nucleic acid and protein sequences. Another CHLOROBX  
55 program is GeSeq, an application for a rapid and accurate annotation of organelle genomes [4].  
56 GenBank 2 Sequin can be used to convert GeSeq’s annotation output for database submission.  
57 Nonetheless, any custom GenBank file can be prepared for NCBI submission using GenBank 2 Sequin.

## 58 **Implementation**

59 GenBank 2 Sequin is a web application written in Java and JavaScript, and tested with the current  
60 versions of the JavaScript enabled browsers Firefox, Chrome, IE11 and Edge. It has implemented tbl2asn  
61 v25.3 (<http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2>) for creation of Sequin files.

62

## 63 **Results**

64 GenBank 2 Sequin parses the GenBank file and converts the annotation into a tab delimited annotation  
65 table (“five-column, tab-delimited feature table”). It further extracts the nucleic acid sequence  
66 information from the GenBank file and writes it, together with the mandatory source and sequence  
67 information of an NCBI record (see below), into a FASTA file. These two files can already be used for  
68 submission through BankIt or to update an existing GenBank record. To create Sequin files for direct  
69 submission, GenBank 2 Sequin invokes tbl2asn. For this, it combines annotation table, FASTA file, and  
70 additional files that contain sequence source or author submission information (see below). As an  
71 optional feature, GenBank 2 Sequin can edit or add gene product names of coding sequences (CDS),  
72 tRNAs, and/or rRNAs in or to the annotation. This might be helpful for the revision of larger genomes.  
73 Last, GenBank 2 Sequin produces several output files for quality control (Fig. 1).

74

### 75 *File upload*

76 Several file types can be uploaded to GenBank 2 Sequin:

77

### 78 *GenBank file*

79 This file is mandatory and must contain the LOCUS information (either an accession number or a user  
80 defined identifier), the sequence FEATURES according to the standards of the International Nucleotide  
81 Sequence Database Collaboration (INSDC, [http://www.insdc.org/documents/feature\\_table.html](http://www.insdc.org/documents/feature_table.html)), and  
82 the ORIGIN, i.e. the nucleic acid sequence in the GenBank format. Currently GenBank 2 Sequin does not  
83 accept multi-GenBank files. Please note that FEATURES, or included qualifiers therein, which are not  
84 concise with the INSDC syntax might be modified or removed by tbl2asn (also see below). All other  
85 entries in the GenBank file, such as submitter’s information, literature references, definition line, or  
86 source information are ignored and must be provided separately by the following files or input options.

87 *Author Submission Template*

88 This file contains submitter's information and literature references, which will be later displayed in the  
89 final database entry. The template can be created at NCBI  
90 (<https://submit.ncbi.nlm.nih.gov/genbank/template/submission>). If no Author Submission Template is  
91 provided, GenBank 2 Sequin will use "Unknown Author" as default. Submitter's information and  
92 literature references are also modifiable later in the Sequin file.

93

94 *Source Modifier Table*

95 This optional upload can contain non-mandatory information for the sequence source description, such  
96 as collection site of an organism, voucher information, or a note. For controlled vocabulary see  
97 <http://www.ncbi.nlm.nih.gov/Sequin/modifiers.html>. The data can be either provided in the source  
98 table format \*.src, (<https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2>), or as a two-column, tab  
99 delimited text (an exemplary file can be download from the GenBank 2 Sequin web interface). Again,  
100 sequence source modifiers can also be added manually to the Sequin file prior to submission.

101

102 *Gene Product Specification Table*

103 This table is optional as well and might be useful to revise and update lager genomes: With the help of a  
104 two column, tab-delimited text file (see example at the GenBank 2 Sequin homepage), GenBank 2  
105 Sequin will either add or change gene product names of the annotation features CDS, tRNA or rRNA. For  
106 instance, if the Gene Product Specification Table contains the line "psaA [tab character] photosystem I  
107 P700 apoprotein A1", GenBank 2 Sequin will search for "psaA" in the annotation. If no gene product  
108 name for "psaA" was provided in the original GenBank file (which, e.g., is the case for GeSeq output)[4],  
109 GenBank 2 Sequin will add "photosystem I P700 apoprotein A1" as gene product name. If in the original  
110 GenBank file the gene product of "psaA" was differently specified, for example as "PSI-A core protein of  
111 photosystem I", this description will be replaced by "photosystem I P700 apoprotein A1". If "psaA" is not  
112 present in the annotation or in the Gene Product Specification Table, no action will be taken.

113

114 *Options*

115 In this window, the user can add/select mandatory source and sequence information, such as source  
116 organism, molecule type, location, genetic code and indicate if the sequence is complete and/or circular.  
117 In addition, the user can specify the definition line of the GenBank record.

118 *Output*

119 GenBank 2 Sequin produces several output files: (i) the nucleic acid sequence in FASTA format with  
120 mandatory sequence information in the FASTA header, (ii) the annotation table, and (iii) the final Sequin  
121 file for direct submission. The first two files can be used for submission through the BankIt web interface  
122 or for the update of an existing GenBank entry.

123 The remaining files are for quality control: (iv) the tbl2asn log file reports any syntax errors in  
124 the original annotation as identified by tbl2asn. Those can include unknown qualifiers or feature names  
125 not allowed by NCBI. Consequently, the tbl2asn program corrects them by removing unknown qualifiers  
126 and/or changing any unknown features into misc\_features (see above). (v) The annotation, as it will be  
127 displayed later in NCBI, is recorded in the GenBank output. Changes in the annotation due to potential  
128 conversion errors and/or modifications can be easily identified by comparing the user's original  
129 GenBank file with GenBank 2 Sequin's GenBank output using the comparison function found in many  
130 common text editors such as Microsoft Word (Microsoft Corporation, Redmond, WA, USA). (vi) Last,  
131 validation of the annotation by tbl2asn ([https://www.ncbi.nlm.nih.gov/genbank/genome\\_validation](https://www.ncbi.nlm.nih.gov/genbank/genome_validation)) is  
132 provided in the files "Validation" and "Validation Summary". Prior to submission identified annotation  
133 errors listed therein should be corrected and warnings checked. The most suitable program to correct  
134 these errors is Sequin, which also allows revalidation. Corrected files can be directly submitted.

135

136 **Conclusion**

137 GenBank 2 Sequin is a web application that converts GenBank format annotations into the NCBI  
138 submission format Sequin. Especially for less computer-savvy wet lab researchers the tool substantially  
139 simplifies sequence submission to NCBI and hence, fills an important gap for the community. With its  
140 annotation validation by default, it also will help to improve the quality of sequence database entries.

141 **Funding**

142 This work has been supported by grants from the Deutsche Forschungsgemeinschaft (DFG; GR 4193/1-1)  
143 and by the Max Planck Society to S.G.

144

145 **Authors' contributions**

146 GenBank 2 Sequin was conceived, designed, written and tested by PL and SG.

147

148 **Acknowledgements**

149 We wish to thank Drs. Michel Tillich and Ralph Bock for critical reading and comments on the  
150 manuscript and an anonymous reviewer for helpful suggestions to a previous version. We are grateful to  
151 the IT Service Team of the Max Planck Institute of Molecular Plant Physiology for excellent technical  
152 assistance.

153 **References**

- 154 1. Submitting Sequences using Specific NCBI Submission Tools. In: The GenBank Submissions  
155 Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011.  
156 <https://www.ncbi.nlm.nih.gov/books/NBK53709/>. Accessed 24 Aug 2017.
- 157 2. Kropinski A. Converting GenBank flat files (gbk) to Sequin (sqn) format. A work around for  
158 gbk2sqn. ResearchGate. 2016; doi: 10.13140/RG.2.1.1931.4964
- 159 3. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, et al. Artemis: sequence  
160 visualization and annotation. *Bioinformatics*. 2000;16:944-945.
- 161 4. Tillich M, Lehwarck P, Ulbricht-Jones ES, Fischer A, Pellizzer T, Bock R, et al. GeSeq - versatile and  
162 accurate annotation of organelle genomes. *Nucleic Acids Res*. 2017;45:W6-W11.



163 **Figures**

The image displays the GenBank 2 Sequin Graphical User Interface, divided into three main sections: Files, Options, and Result.

- Files:** This section contains four file upload categories, each with a 'Set File' button and a list of uploaded files:
  - GenBank (required):** 1 file in list: 'My Custom GenBank File.gb'.
  - Author Submission Template (optional)<sup>5</sup>:** 1 file in list: 'My Submission Data.sbt'.
  - Source Modifier Table (optional)<sup>8</sup>:** 1 file in list: 'My Source Modifiers.src'.
  - Gene Product Specification Table (optional)<sup>5</sup>:** 1 file in list: 'My Gene Products.txt'.
- Options:** This section is titled 'Specify your Sequence - NCBI Standard Parameters as Default' and includes:
  - Text input for 'Definition' (My Sequence).
  - Text input for 'Organism' (My Organism).
  - Dropdown for 'Molecule Type' (genomic DNA).
  - Checkboxes for 'Complete' and 'Circular' (both checked).
  - Dropdown for 'Genetic Code' (Standard).
  - Dropdown for 'Location (e.g. Organelle)' (genomic).
  - 'Start Conversion' and 'Reset' buttons.
  - Checkbox for 'I have read and accept the Disclaimer' (checked).
- Result:** This section displays the output files:
  - FASTA
  - Annotation Table
  - tbl2asn log
  - Sequin
  - GenBank
  - Validation
  - Validation Summary
  - 'Clear Results' button.

164  
165 **Fig. 1** GenBank 2 Sequin Graphical User Interface. Left column (Files): File upload for the mandatory  
166 GenBank file, and optional files such as Authors Submission Template, Source Modifier Table and Gene  
167 Product Specification Table. Right column, upper box (Options): Window to specify the definition line of  
168 the GenBank entry and to add mandatory source and sequence information, such as organism or  
169 molecule type. Right column, lower box (Results): GenBank 2 Sequin output: FASTA file, annotation  
170 table, tbl2asn log, and final Sequin file for submission. For the user's review the GenBank entry as it will  
171 be displayed later in NCBI is provided, as well as two files that include the validation of the annotation.  
172 Files can be downloaded by pressing the respective buttons. For details, see main text.