

Adding Vectors to Sequencher's VecBase File

In these pages we first describe the essential features of the CODATA format, which is the standard format used to store vector information. We then describe the steps necessary for adding your own vector information to Sequencher's VecBase file.

VecBase File Format: Description of Essential Features

Vecbase is stored in CODATA format. It was defined by the CODATA Task Group on Coordination of Protein Sequence Data Banks. The full description is in an obscure reference:

Protein Seq Data Anal (1987)1:27-39

As we describe the general specifications of the CODATA format you can see examples in the *VecBase* file itself if you open it in a general word processing program, such as Microsoft Word, or a plain text editor.

1. The '\' character (backslash) is special. It is used as a general data item separator and indicates that what follows is a separate data item but of the same type. The lines immediately preceding and immediately following the main body of data must have only three backslash characters "\\\".
2. Different types of data within a single database entry are distinguished by dividing them into specific data items, e.g., title, reference, feature table, etc. Space characters are used as general separators to separate data. Each data item is labeled with an *Identifier*. *Identifiers* are single words or several words connected by hyphens.

Note:

- The first three characters of each identifier must be unique.
 - *Identifiers* cannot contain space characters.
 - A data item may extend over as many lines as necessary.
 - The identifier **MUST** start at the first column of the first line corresponding to the item.
 - Continuation lines are distinguished by containing at least three space characters at the beginning of the line.
3. Data within data items are divided into fields (subitems). Each field consists of a *subidentifier*, which identifies the field and separates subitems, followed by the associated data. Subidentifiers are of the same form as identifiers, but must be immediately preceded by a number sign, '#'

4. An entry can consist of the following (non-exhaustive) list of data items:

Beginning-of-entry
Title
Alternate-name
Includes
Date
Accession
Source
Host
Reference
Comment
Genetic
Keywords
Cross-reference
Feature
Physical
Origin
Summary
Sequence
End-of-entry

- *Beginning-of-entry* is always the word "**ENTRY**"
- *End-of-entry* is always three slash characters, "///".
- *Title* always comes immediately after *Beginning-of-entry*;
- *Summary* and *Sequence* data always come just before *End-of-entry*.

5. Immediately after the **ENTRY** keyword comes the *Entry Identification Code*. This is a unique, one- to ten-character word containing no spaces.. The *Entry Identification Code* is followed by a subitem specifying the type of molecule (DNA, cDNA, Protein, tRNA, etc.) denoted by the subidentifier **#Type**.

example:

Entry ZEBPAL #Type Protein Fragment

Entry OKBOG #Type PROTEIN

Entry HUMIT #DNA circular, double stranded

6. You should not have more than 80 characters per line according to the CODATA specification.

That's some of the general format of *VecBase*, but we only need to worry about a few fields in order to work in Sequencher:

Adding New Vectors to Sequencher's VecBase file

1. Open the VecBase file in any word processing or file editing program. It should be in the Sequencher folder on your hard drive.
2. The easiest way to add a new vector is to first duplicate the complete record of an existing vector in VecBase, and edit the duplicate with the correct information for the new vector. (Remember to copy through the end-of-entry symbol “\\”).
3. Edit the following three fields with the correct information for the new vector:

ENTRY {name} where {name} is replaced by the name of the vector as it will appear in the vector selection list.

POLYLINKER - list the set of sites in the polylinker, separated by hyphens, e.g.:

e.g. KpnI-DraII-ApaI-XhoI-SalI-ClaI-HindIII-EcoRV-EcoRI-PstI-SmaI-BamHI-SpeI-XbaI-NotI-XmaIII-BstXI-SacII-SacI

SEQUENCE - numbers at the beginning of each line and blocking in groups of 10, as shown here, are optional, but remember you need at least 3 blank spaces at the beginning of each continuation line for an item.

Example (the numbers and spaces are here only for readability and are not necessary):

```

1  GGACGCGCCC   TGTAGCGGCG   CATTAAAGCGC   GCGGGGTGTG   GTGGTTACGC
51 GCAGCGTGAC   CGCTACACTT   GCCAGCGCCC   TAGCGCCCGC   TCCTTTCGCT
101 TTCTTCCCTT   CCTTCTCGC   CACGTTCCGCC   GGCTTTCGCC   GTCAAGCTCT
151 AAATCGGGGG   CTCCTTTAG   GGTTCCGATT   TAGTGCTTTA   CGGCACCTCG
201 ACCCAA AAAA   ACTTGATTAG   GGTGATGGTT   CACGTAGTGG   GCCATCGCCC
251 TGATAGACGG   TTTTTC...
```

4. This is followed by the End-of-entry symbol, '\\'
5. Save the file under the name “VecBase” as a **text-only** file.
6. When you next launch Sequencher and choose **Select Vector Insertion Site** from the **Windows** menu, the new vector should appear in the alphabetical listing.