User Documentation for `ENCprime`

John Novembre

February 28, 2006

Introduction

Included in this package are two programs that can be used to calculate a set of codon usage bias summary statistics. One of these statistics is $\hat{N}_c'$, a statistic I developed that can account for background nucleotide composition. The publication describing this statistic is "Accounting for background nucleotide composition" *Molecular Biology and Evolution* 19(8):1390-1394. 2002.

The two programs included in this distribution are:

- `SeqCount`: This program will take FASTA formatted sequences and produce tables that describe the numbers of each codon/nucleotide observed in those sequences as well as tables for the frequencies of each codon/nucleotide observed.
- `ENCprime`: This program will take the tables output by `SeqCount` and calculate various codon usage bias summary statistics for each sequence. The output of `ENCprime` is a file with a table describing the results. There is also an optional interactive function of `ENCprime` that allows you to peruse the data and results of the analysis.

For those who only need an example to get off and running, here is an example of using these programs from the command-line (Win/Unix only):

```
SeqCount -c ExSeqs.fasta 9
SeqCount -n ExBackgroundSeqs.fasta 9
ENCprime ExSeqs.fasta.codcnt ExBackgroundSeqs.fasta.acgtfreq 1 ExResults 2
```

The first two lines translate the FASTA input into the table formats that ENCprime can read. The 9 in both lines is because there are 9 sequences in the fasta files. The third line runs ENCprime on the two translated files with the standard genetic code (hence the 1), outputs the results to a file `ExResults`, and does so with high verbosity (lots of info spews on the screen, this is set by the 2).

The next two sections provide more detail about what's going on and additional options one can use.

Using SeqCount

1. **Standard interface:** Run `SeqCount`. The program will prompt you for the name of the FASTA formatted set of sequences. Then it will prompt you whether codons or nucleotides should be counted, and how many sequences your FASTA file contains.

2. **Using command line parameters:** Simply type: `SeqCount <-c/-n> <filename> <Num Seqs>` . Use `-c` to produce a codon count file, and `-n` to produce a nucleotide composition file. `<filename>` should be the name of the FASTA formatted input file, and `<Num Seqs>` should be the number of sequences in that file.

If counting codons, the program will output two files. One file will have the suffix `.codcnt` and it will contain the raw counts of each codon used in each sequence. The other file will have the suffix `.codfreq` and it will contain the frequency of each codon in each sequence.

If counting nucleotides, the program will output two files that describe the nucleotide composition. The file with suffix `.acgtcnt` will contain the raw counts of each nucleotide in the sequence as well as the length of the sequence. The second file will have the suffix `.acgtfreq` and will contain the frequencies of each nucleotide in each sequence.

For example, if we were counting codons with input file `yeast_sequences`, the output would be `yeast_sequences.codcnt` and `yeast_sequences.codfreq`. If counting nucleotides, the program will output the files `yeast_sequences.acgtcnt` and `yeast_sequences.acgtfreq`.

When we go on to use `ENCprime`, we will use the files with suffixes `.codcnt` and `.acgtfreq`. The additional files produced by `SeqCount` are not used by `ENCprime`, but they may be useful for other analyses.

Viewing data across all sequences: For viewing data from across all the sequences in a dataset, the program outputs an extra line labeled "Totals>" in each output file. This line is formatted just like that of the rest of the output so that when `ENCprime` is run on the `SeqCount` output you will be able to obtain codon bias statistics based on all the sequences of a data set taken together.

Running ENCprime

Like with `SeqCount` there are two ways to run the program. If you type simply `ENCprime` you will enter a menu-driven mode of selecting the options. You can also set the options using command-line parameters (see below).

Options:

- Codon counts file: This is a file containing the counts of each codon. See `ExCodCount.codcnt`. These files can be produced from FASTA sequences by the `SeqCount` program that I've included (see above).

- Nucleotide composition file. This file describes the background nucleotide composition for each of the sequences. There are two ways to input your data:

  - The first is to give the background nucleotide compositions in terms of frequencies. In this case, each line of the file is the name of the sequence and four numbers corresponding to the background frequency of A,C,G,T (in that order). The expected frequency of each codon will be computed by the program using these nucleotide frequencies. The method to do this is described in the MBE paper. To alert the program you're using this format, the first word of the first line of the file should start with the letter 'N'. For an example, see `ExNucComp.acgtfreq`. These files can be produced with `SeqCount` (see above).
  - The second, more general method is to input the expected frequency of each codon. Here each line of the file will have the sequences name and then 64 numbers corresponding to the expected frequencies of each codon's usage. Note that the expected frequencies for a set of synonymous codons should add to 1. To alert the program you're using this format, the first word of the first line of the file should start with the letter 'C' (see `ExNucComp.excodfreq`, for an example). When using this option, the codons should be input in the same order as in the example file.

  Note: `ENCprime` will choke on input files where the sequence name contains the character `>`. This occurs because `ENCprime` uses the `>` as the delimiter between sequence name and codon usage data.

- Genetic code. This option describes the genetic code to be used in translating the sequences. The genetic codes provided by the NCBI on the Taxonomy browser (`http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c`) are available by entering the corresponding translation table ID for this option (for example, id 5 is the invertebrate mitochondrial code). Otherwise, you can use a custom genetic code by entering a filename. For an example of the file format see `ExGeneticCode.dat`.

- Output file name. This will be the name of the output file generated by the program.

- Verbosity. When set to 1, the program will output to the screen all the data as it reads it in. This is useful for checking your input files are formatted correctly. When set to 2, it will output the data but with pauses to allow you to look over the input. Note: setting the verbosity to 2 is a good way to make sure the program is reading in your data correctly. If your data is not being read correctly, check the formatting of the relevant input file.

- Data Explorer: The `Data Explorer` is a fairly straightforward way of viewing all of the results of the analysis. Setting this option to 0 will have you enter data explorer after the analysis, while 1 will have you skip it.

Finally, a note on the default values. These are controlled by a file called `ENCprime.defaults` and by editing this file you can change the defaults.

Command-line mode

Depending on the operating system, you can also set the options from the command-line. This is useful for incorporating the program in to shell scripts. An example is provided in the introduction (see above). Here are the command-line parameters in the order they should appear:

```
ENCprime <count file> <nuc comp file> <genetic code> <output file>
<verbosity> <-q>
```

Note: Typing `-q` as the last parameter will cause the program to skip the `Data Explorer`. This is useful because the Data Explorer requires keyboard input, which can be a problem if you're building scripts based on `ENCprime`.

Output

The program will output one file per run. The name of the file is an option set in the menus or the command-line. The file contains the summary statistics for each sequence in a space-delimited table with labeled columns.

The statistics output per sequence are:

1. Nc, the effective number of codons, $\hat{N}_c$, originally described by Wright 1990.
2. Ncp,the effective number of codons prime, $\hat{N}'_c$, described in Novembre 2002.
3. ScaledChi, Akashi's scaled χ^2.
4. SumChi, the sum of the chi-square statistics, unscaled.
5. df, the corresponding number of degrees of freedom for that statistic
6. p, the p-value for that chi-square statistic
7. B_KM, the $B^*(a)$ measure of Karlin and Mrazek.
8. n_codons, the number of codons for that sequence.

Short sequences and comparing `ENCprime` to other programs

Some users have noted that values of $\hat{N}_c$ calculated by `ENCprime` differ from those calculated by other programs, particularly for shorter coding sequences. The main reason for this is that algorithms differ in how they handle the case where few codons of any particular amino acid are observed.

The following paragraph of the original paper describes in part the approach used in `ENCprime` to handle short sequences:

> In the practical implementation of both of $\hat{N}_c$ and $\hat{N}'_c$, care must be taken to exclude $\hat{F}_a$ values that are undefined or equal to zero, as occurs when amino acids are rare or missing (see Wright 1990). The calculations presented below follow Wright's suggestion that if no 3-fold redundant codons are observed, one should average $\bar{\hat{F}}_2$ and $\bar{\hat{F}}_4$ to obtain $\bar{\hat{F}}_3$. If other k-fold redundancy classes are unobserved $\bar{\hat{F}}_k$ is assumed to equal $1/k$. Such an assumption is conservative with regards to measuring strong codon usage bias. Finally, in the calculation of $\bar{\hat{F}}_k$ we exclude amino acids that are observed fewer than five times. Here these methods of dealing with missing data are applied equally to both the calculation of $\hat{N}_c$ and $\hat{N}'_c$.

There is one typo in the above paragraph. The second to last sentence should read: In the calculation of $\bar{\hat{F}}_k$ we exclude amino amino acids that are observed *five or fewer* times.

There is also a correction not mentioned in the above paragraph. If for any amino acid $\hat{F}_a$ is less than $1/k$ we correct it to $1/k$. This correction is in line with Wright's suggestion on page 25 of his publication that if the observed useage is more uniform than expected by chance the value of $\hat{N}_c$ should be revised. The approach taken here though applies the correction at the level of each amino acid.

In general these corrections only affect the behavior of `ENCprime` when short sequences are used. They are useful heuristics intended to make $\hat{N}_c$ and $\hat{N}'_c$ conservative measures of codon bias when sequence lengths are short and information-poor. Other programs may use different correction procedures. For datasets with large sequences, users should probably not be too concerned regarding which program they use to compute $\hat{N}_c$, but for datasets with numerous short sequences, users should use the algorithm with the correction procedures that they deem most appropriate for their purposes.

Some final remarks

Please feel free to e-mail me (`novembre@berkeley.edu`) if you are having trouble with this software or if you believe you've found a bug. Before you do so though, please be sure to run the program with verbosity set to 2 and check to make sure the input is being read in correctly. Most of the times the program fails for me is it because I have made a mistake in the input.

Software license

IN NO EVENT SHALL REGENTS BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, INCLUDING LOST PROFITS, ARISING OUT OF THE USE OF THIS SOFTWARE AND ITS DOCUMENTATION, EVEN IF REGENTS HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

REGENTS SPECIFICALLY DISCLAIMS ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE SOFTWARE AND ACCOMPANYING DOCUMENTATION, IF ANY, PROVIDED HEREUNDER IS PROVIDED "AS IS". REGENTS HAS NO OBLIGATION TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.