SSUalign User's Guide

Structural alignment of small subunit ribosomal RNA sequences

http://infernal.janelia.org/
Version @RELEASE@; @RELEASEDATE@

Eric Nawrocki and Sean Eddy
HHMI Janelia Farm
19700 Helix Drive
Ashburn VA 20147
http://selab.janelia.org/

Contents

1 Introduction

SSU-ALIGN identifies and aligns small subunit ribosomal RNA (SSU rRNA) genes in sequence datasets. It uses a model of the primary sequence and secondary structure conservation of SSU rRNA as inferred by comparative sequence analysis and confirmed by crystal structure from the Comparative RNA Website (CRW).

Overview

SSU-ALIGN uses profile probabilistic models for SSU rRNA detection and alignment. It includes five SSU rRNA models, an archaeal model, a bacterial model, a eukaryotic (nuclear) model, a chloroplast model and a animal mitochondrial model. These are the default models used by the program, but SSU-ALIGN can be used with other, user-generated models as well.

Given an input dataset of unaligned SSU sequences, SSU-ALIGN goes through two stages. In stage 1, each of the sequences is locally aligned to each of the five models based on primary sequence conservation to obtain a score for each sequence to each model. The model that gives the highest scoring alignment to each sequence is the *best-matching model* for that sequence. In stage 2, each sequence is aligned to it's best-matching model using both primary sequence and secondary structure conservation. The end result is a separate multiple alignment for each model that was the best-matching model for at least one sequence in the input dataset.

The basic tutorial in section 3 of this guide walks through this two-stage process for a simple dataset.

The stage 1 primary sequence alignment is local with respect to both the model and the sequence. Locality with respect to the model allows the program to handle sequences that are subsequences of SSU rRNA that correspond to a specific region of the molecule. Locality with respect to the sequence allows the program to handle input sequences that contain extra residues that are nonhomologous to SSU at the beginning and/or end of the input sequence. In this case, the local alignment will only correspond to a subsequence of the input sequence, it is only this subsequence that is aligned to it's best-matching model in stage 2.

The stage 2 primary sequence and structural based alignment is local *only* with respect to the model. This means the subsequence that survives stage 1 is always included in it's entirety in the final alignment.

This package can also be used to remove, or prune, unreliable alignment columns from the final alignments. This post-processing step can be useful for removing regions of the alignment that are are ambiguous and likely include a significant number of errors prior to using phylogenetic inference tools.

Columns are selected for pruning based on "confidence estimates" for the aligned residues, which are derived from the probability that each residue aligns to each column of the alignment. This is discussed in more detail in the basic tutorial in section 3.

This package can also be used to create new SSU models, either for different phylogenetic clades, or for specific regions of the SSU rRNA molecule, as well as for drawing secondary structure diagrams that display several different statistics on the alignment (such as levels of sequence conservation, or frequency of gaps). The advanced tutorials in section X-X of this guide provide examples of these applications.

What is included in this package

This distribution includes the PERL script SSU-ALIGN and all the files necessary to compile and run the INFERNAL package (version 1.0) (Nawrocki et al., 2009). It also includes the five default (archaea, bacteria, chloroplast, eukarya, and animal-mitochondria) SSU models and the *seed* alignments they were built from.

The alignments were based on SSU structures and alignments from the Comparative RNA Website (CRW) (**?**). A broader discussion of these models and alignments and the specific procedure used to create the seed alignments from the CRW data is explained in section 2.

Installation of the PERL (Practical Extraction and Report Language, Larry Wall) interpreter package version 5.0 or later is required to run the SSU-ALIGN PERL script.

What *SSU-align* does not do

SSU-ALIGN only creates alignments it does not infer trees from those alignments. SSU-ALIGN also does not classify sequences beyond reporting which model in the input CM file they score highest to.

Future development

This release (version 0.1) is a prototype of the SSU-ALIGN software package that only has basic functionality for aligning SSU sequences. I hope to continue to develop and improve it in the future by making it faster and generally more useful to the SSU rRNA sequence analysis community. I welcome bug reports as well as feature requests.

2 The SSU ribosomal RNA secondary structure models used by SSU-ALIGN

The profile probabilistic models of SSU rRNA in SSU-ALIGN are based on the alignments generated by Robin Gutell and colleagues at the University of Texas Austin (CITE). There are 5 CM files, each was constructed from a separate alignment from the Comparative RNA Website. Statistics on those alignments are in Table X.

Procedure for going from CRW to consensus structure annotated Stockholm alignment files

Training alignment statistics mapped onto consensus secondary structures

DIAGRAMS OF PRIMARY SEQUENCE CONSERVATION, STRUCTURE INFO, DELETIONS, INSERTIONS, STARTING POINTS.

3 Basic tutorial: defining and aligning SSU sequences using SSU-ALIGN

Here is a tutorial walk-through of a small project with SSUALIGN. This tutorial shows how to use the program for it's most basic and fundamental purpose, creating multiple alignments of SSU rRNA sequences.

Files used in this tutorial

The subdirectory **/tutorial-basic** in the SSU-ALIGN contains the files used in this tutorial, they are:

ssu.default.0p1.cm A covariance model (CM) file that defines five SSU rRNA CMs: an archael model, a bacterial model, a choloroplast model, a eukaryotic model and a metazoan mitrochondria model. These are the five default models used by SSU-ALIGN. These models are explained in section 4.

rocks.fa SSU rRNA sequences from an environmental survey sequencing project of microbes living in the pore space of rocks in the Rocky Mountains by J.J. Walker and Norm Pace (Walker and Pace, 2007).

1p0.params A file containing paths to INFERNAL executable files that SSU-ALIGN needs to run. You will likely need to change these paths to point to where you've installed the **cmsearch** and **cmalign** programs (these are created in 'infernal-1.0/src/' after building INFERNAL version 1.0 with 'sh ./configure; make;') and the **esl-sfetch** program (which is created in 'infernal-1.0/easel/miniapps/' after building INFERNAL version 1.0).

Create a new directory that you can work in, and copy all the files in **tutorial-basic** there. I'll assume for the following examples that you've installed the SSU-ALIGN script in your path; if not, you'll need to give a complete path name to the script (e.g. something like **/usr/people/nawrocki/ssualign/src/ssu-align** instead of just **ssu-align**).

The file **rocks.fa** contains 588 SSU sequences (Walker and Pace, 2007). SSU-ALIGN is designed to create structural alignments of SSU sequences from studies like this one. To run it, execute the following command:

```
> ssu-align ssu.default.0p1.cm rocks.fa myrun 1p0.params
```

The program will report on what its doing:

```
#
# Stage 1: Defining SSU start/ends with cmsearch (crude time estimate: 0.3 minutes) ...
```

In stage 1, the program scans the input sequences with each of the three models in the CM file **abcem.cm**. This has two purposes. First, it classifies each sequence by determining which model gives each sequence the highest HMM alignment score. Secondly, it defines the start and end points of the SSU sequences.

When this step finishes, you'll see:

```
#
# Stage 1: Defining SSU start/ends with cmsearch (crude time estimate: 1.1 minutes) ...  done.
#
# output file name                                                  description
# ----------------------------------------------------------------  -----------
  myrun/myrun.tab                                                   cmsearch tabular output file with locations/
  myrun/myrun.archaea.sseq.list                                     list of high scoring subseqs to align with a
  myrun/myrun.archaea.sseq.fa                                       FASTA file of high scoring subseqs to align
  myrun/myrun.bacteria.sseq.list                                    list of high scoring subseqs to align with b
  myrun/myrun.bacteria.sseq.fa                                      FASTA file of high scoring subseqs to align
  myrun/myrun.cholorplast.sseq.list                                 list of high scoring subseqs to align with c
  myrun/myrun.cholorplast.sseq.fa                                   FASTA file of high scoring subseqs to align
```

This lists and briefly describes the 7 new files the script created in a subdirectory called **myrun/**. The archaeal model was the best match to 48 of the 588 sequences. The **myrun/myrun.archaea.sseq.list** file lists these sequences. The **myrun/myrun.archaea.sseq.fa** contains the 48 subsequences. 341 sequences were scored highest by the bacterial model. The list and sequences are in **myrun/myrun.bacteria.sseq.list** and **myrun/myrun.bacteria.sseq.fa**. The remaining 199 sequences were scored highest by the chloroplast model, these are listed in **myrun/myrun.chloroplast.sseq.list**; the actual sequences are in **myrun/myrun.chloroplast.sseq.fa**.

There were no sequences that best-matched the eukaryotic model or the metazoan mitochondria model.

The output of INFERNAL'S **cmsearch** program is in the file **myrun/myrun.tab**.

The program will now proceed to stage 2, the alignment stage. This stage serial progresses through each model that had at least 1 matching sequence and aligns the sequences to the model using both structure and sequence conservation. A time estimate is provided for each stage.

```
#
# Stage 2. Aligning sequences.
#
# stage  cm                   seq file                                                    nseq  est m
# -----  -------------------  ----------------------------------------------------------  -------  -----
   1/ 5  archaea              0615-1/0615-1.archaea.sseq.fa                                 48     0.
   2/ 5  bacteria             0615-1/0615-1.bacteria.sseq.fa                               341     6.
   3/ 5  cholorplast          0615-1/0615-1.cholorplast.sseq.fa                            199     3.
   4/ 5  eukarya              NONE (no matching seqs)                                        0     0.
   5/ 5  mitochondria-animal  NONE (no matching seqs)                                        0     0.
```

After the alignment stage ends there will be three new alignment files: **myrun/myrun.m1.cmalign.stk** and **myrun/myrun.m2.cmalign.stk**.

Take a look at the archaeal alignment in **myrun/myrun.m1.cmalign.stk**.

This alignment includes consensus secondary structure annotation and is in *Stockholm format*. Stockholm format, the native alignment format used by HMMER and INFERNAL and the PFAM and RFAM databases, is documented in detail in the INFERNAL User's Guide which is included in this distribution in **infernal/documentation/userguide.pdf**.

For now, what you need to know about the key features of the input file is:

- The alignment is in an interleaved format, like other common alignment file formats such as CLUSTALW. Lines consist of a name, followed by an aligned sequence; long alignments are split into blocks separated by blank lines.
- Each sequence must have a unique name that has zero spaces in it. (This is important!)
- For residues, any one-letter IUPAC nucleotide code is accepted, including ambiguous nucleotides. Case is ignored; residues may be either upper or lower case.
- Gaps are indicated by the characters ., _, -, or ~. (Blank space is not allowed.)
- A special line starting with `#=GC SS_cons` indicates the secondary structure consensus. Gap characters annotate unpaired (single-stranded) columns. Base pairs are indicated by any of the following pairs: `<>`, `()`, `[]`, or `[]`. No pseudoknots are allowed; the open/close-brackets notation is only unambiguous for strictly nested base-pairing interactions.
- The file begins with the special tag line `# STOCKHOLM 1.0`, and ends with `//`.

To convert the alignment to fasta format that includes gaps, you can use the **`scripts/stk2aln_fa.pl`** script.

NOT SURE WHAT TO WRITE ABOUT THE ALIGNMENT!

Pruning the alignment based on probabilistic confidence estimates

4 Tutorial

Here's a tutorial walk-through of some small projects with SSUALIGN. This section should be sufficient to get you started on work of your own, and you can (at least temporarily) skip the rest of the Guide.

Files used in the tutorial

The subdirectory **`/tutorial`** in the SSUALIGN distribution contains the files used in the tutorial, as well as a number of examples of various file formats that SSUALIGN reads. The important files for the tutorial are:

`abcem.cm` A covariance model (CM) file that defines five SSU rRNA CMs: an archael model, a bacterial model, a choloroplast wmodel, a eukaryotic model and a metazoan mitrochondria model. These are the five default models used by textscSSUalign.

`rocks.fa` SSU rRNA sequences from the an environmental survey sequencing project of microbes living in the pore space of rocks in the Rocky Mountains by J.J. Walker and Norm Pace (Walker and Pace, 2007).

Create a new directory that you can work in, and copy all the files in **`tutorial`** there. I'll assume for the following examples that you've installed the SSUALIGN programs in your path; if not, you'll need to give a complete path name to the programs (e.g. something like **`/usr/people/nawrocki/ssualign/src/ssu-align`** instead of just **`ssu-align`**).

Defining and aligning SSU sequences using SSUalign

The file **`rocks.fa`** contains 588 SSU sequences obtained from an environmental survey of the pore space of rocks in the Rocky Mountains by J.J. Walker and Norm Pace (Walker and Pace, 2007). SSUALIGN is designed to create structural alignments of SSU sequences from studies like this one. To run it, execute the following command:

```
> ssu-align abcem.cm rocks.fa myrun 1p0.params
```

The program will report on what its doing:

```
#
# Stage 1: Defining SSU start/ends with cmsearch (crude time estimate: 0.3 minutes) ...
```

In stage 1, the program scans the input sequences with each of the three models in the CM file **`abcem.cm`**. This has two purposes. First, it classifies each sequence by determining which model gives each sequence the highest HMM alignment score. Secondly, it defines the start and end points of the SSU sequences.

When this step finishes, you'll see:

```
#
# Stage 1: Defining SSU start/ends with cmsearch (crude time estimate: 1.1 minutes) ...  done.
#
# output file name                                                  description
# ----------------------------------------------------------  -----------
  myrun/myrun.tab                                             cmsearch tabular output file with locations/
  myrun/myrun.archaea.sseq.list                               list of high scoring subseqs to align with a
  myrun/myrun.archaea.sseq.fa                                 FASTA file of high scoring subseqs to align
  myrun/myrun.bacteria.sseq.list                              list of high scoring subseqs to align with b
  myrun/myrun.bacteria.sseq.fa                                FASTA file of high scoring subseqs to align
  myrun/myrun.cholorplast.sseq.list                           list of high scoring subseqs to align with c
  myrun/myrun.cholorplast.sseq.fa                             FASTA file of high scoring subseqs to align
```

This lists and briefly describes the 7 new files the script created in a subdirectory called **myrun/**. The archaeal model was the best match to 48 of the 588 sequences. The **myrun/myrun.archaea.sseq.list** file lists these sequences. The **myrun/myrun.archaea.sseq.fa** contains the 48 subsequences. 341 sequences were scored highest by the bacterial model. The list and sequences are in **myrun/myrun.bacteria.sseq.list** and **myrun/myrun.bacteria.sseq.fa**. The remaining 199 sequences were scored highest by the chloroplast model, these are listed in **myrun/myrun.chloroplast.sseq.list**; the actual sequences are in **myrun/myrun.chloroplast.sseq.fa**.

There were no sequences that best-matched the eukaryotic model or the metazoan mitochondria model.

The output of INFERNAL'S **cmsearch** program is in the file **myrun/myrun.tab**.

The program will now proceed to stage 2, the alignment stage. This stage serial progresses through each model that had at least 1 matching sequence and aligns the sequences to the model using both structure and sequence conservation. A time estimate is provided for each stage.

```
#
# Stage 2. Aligning sequences.
#
# stage  cm                   seq file                                                    nseq  est m
# -----  -------------------  ----------------------------------------------------------  -----  -----
   1/ 5  archaea              0615-1/0615-1.archaea.sseq.fa                                 48     0.
   2/ 5  bacteria             0615-1/0615-1.bacteria.sseq.fa                               341     6.
   3/ 5  cholorplast          0615-1/0615-1.cholorplast.sseq.fa                            199     3.
   4/ 5  eukarya              NONE (no matching seqs)                                        0     0.
   5/ 5  mitochondria-animal  NONE (no matching seqs)                                        0     0.
```

After the alignment stage ends there will be three new alignment files: **myrun/myrun.m1.cmalign.stk** and **myrun/myrun.m2.cmalign.stk**.

Take a look at the archaeal alignment in **myrun/myrun.m1.cmalign.stk**.

This alignment includes consensus secondary structure annotation and is in *Stockholm format*. Stockholm format, the native alignment format used by HMMER and INFERNAL and the PFAM and RFAM databases, is documented in detail in the INFERNAL User's Guide which is included in this distribution in **infernal/documentation/userguide.pdf**.

For now, what you need to know about the key features of the input file is:

- The alignment is in an interleaved format, like other common alignment file formats such as CLUSTALW. Lines consist of a name, followed by an aligned sequence; long alignments are split into blocks separated by blank lines.
- Each sequence must have a unique name that has zero spaces in it. (This is important!)
- For residues, any one-letter IUPAC nucleotide code is accepted, including ambiguous nucleotides. Case is ignored; residues may be either upper or lower case.
- Gaps are indicated by the characters ., _, -, or ~. (Blank space is not allowed.)
- A special line starting with `#=GC SS_cons` indicates the secondary structure consensus. Gap characters annotate unpaired (single-stranded) columns. Base pairs are indicated by any of the following pairs: `<>`, `()`, `[]`, or `[]`. No pseudoknots are allowed; the open/close-brackets notation is only unambiguous for strictly nested base-pairing interactions.
- The file begins with the special tag line `# STOCKHOLM 1.0`, and ends with `//`.

To convert the alignment to fasta format that includes gaps, you can use the **scripts/stk2aln_fa.pl** script.

NOT SURE WHAT TO WRITE ABOUT THE ALIGNMENT!

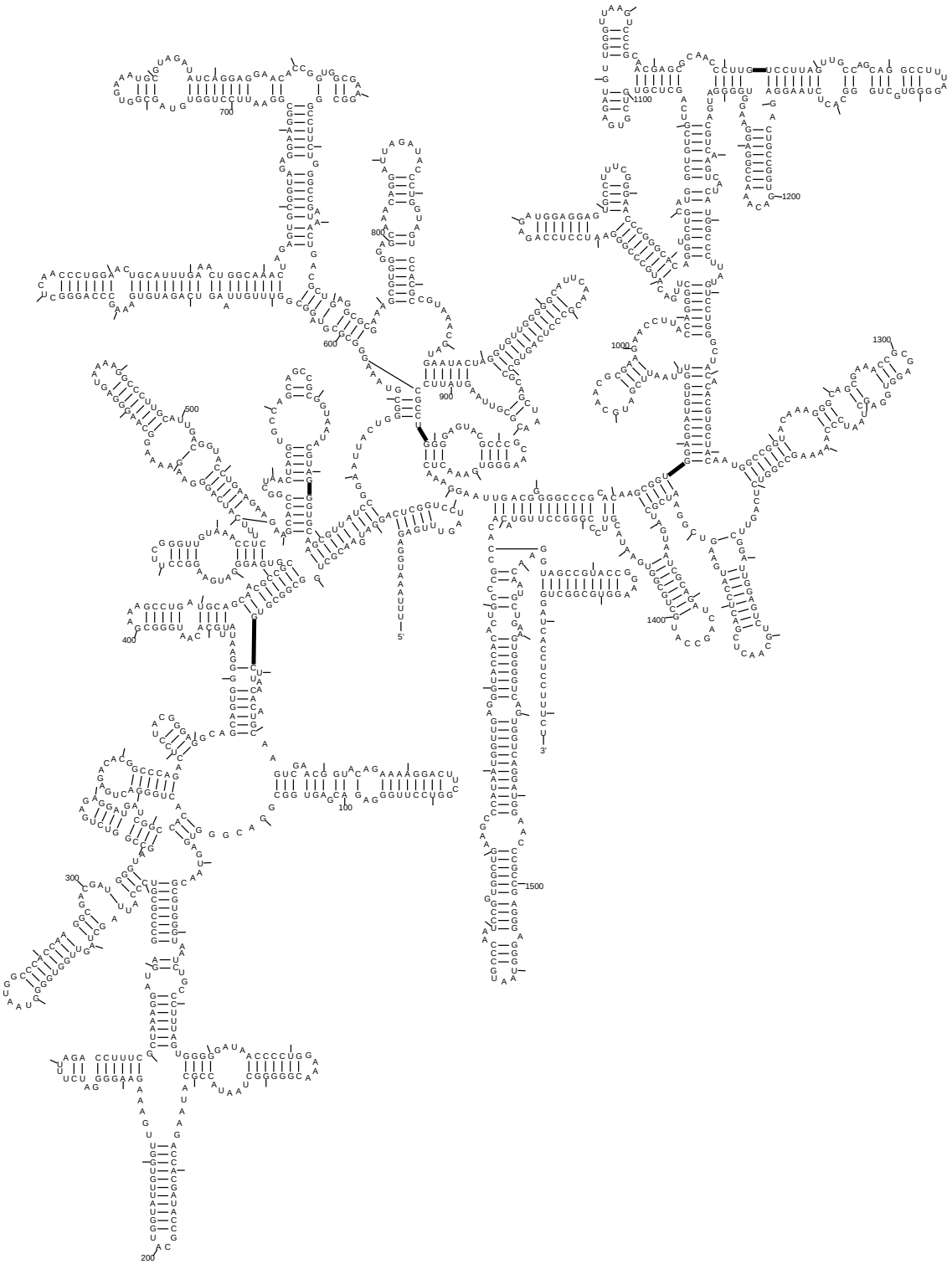
Pruning the alignment based on probabilistic confidence estimates

Creating a truncated model of a specific region of SSU rRNA

Some SSU rRNA sequencing studies target a specific region of the SSU rRNA molecule using PCR primers at the boundaries of that region. For such studies it is recommended to build a new CM that only models the region of the molecule targetted by the study. In this section I'll demonstrate how to create such a CM.

For this example imagine our study is only targetting bacterial SSU rRNA. We will use the bacterial SSU seed alignment that is included with SSUALIGN as a starting point for creating our new, truncated CM. The first step is to determine what consensus positions in the bacterial seed alignment's consensus structure the targetted oregion corresponds to. The consensus structure of the bacterial seed alignment is shown on the next page.

Bacterial small subunit ribosomal RNA (SSU rRNA)



In practice, you'll have to manually find your primer site and determine their positiosn on this consensus structure. Every hundredth residue is numbered, and every tenth residue is marked with a tick mark, which should help.

Imagine for this example our 5' primer begins at consensus position 35 and ends at position 397. The next step is to create a truncated bacterial seed alignment that only models between consensus positions 35 and 397. We can do this with the **esl-alimanip** program using the full bacterial seed alignment as input:

```
> esl-alimanip --start-rf 35 --end-rf 397 -o mybac.stk bacteria.0p1.stk
```

This command creates a new alignment called **mybac.stk** includes the subset of the columns from **bacteria.stk** that lie between consensus positions 35 and 397 inclusively.

The next step is to build a new model from this new alignment with INFERNAL'S **cmbuild** program:

```
> cmbuild --enone --gapthresh 0.8 mybac.cm mybac.stk
```

Note: the **--enone** and **--gapthresh 0.8** flags are recommended best practice for SSUALIGN. The **--enone** flag tells the program to turn off entropy-weighting, a parameterization technique used to make CM homology search more sensitive (**?**) but that seems slightly detrimental to CM alignment accuracy with SSU rRNA models. The **--gapthresh 0.8** flag tells the program to define any column that has less than 80% gaps in the seed as a consensus column. Different values than 0.8 could be used here, but 0.8 seems to yield good performance for SSU alignment, and it was also used to build the five default SSUALIGN models.

Now you can begin using your new model **mybac.cm** to align SSU sequences with SSUALIGN. You have two options.

1. Use **mybac.cm** as the CM file when running SSUALIGN. This is recommended if you believe all of your SSU sequences are bacterial SSU sequences within the specified region (consensus positions 35 to 397).

2. Use **mybac.cm** as one of several models in a multi-model CM file when running SSUALIGN. You can create multi-model CM files by simply concatenating them together. For example you can add it as a sixth model to the default SSUALIGN 0.1 five model CM file **sa.01.abcem.cm** with: **> cat sa.01.abcem.cm mybac.cm > six.cm**. This is recommended if you think only some of your sequences will be within the specified region (consensus positions 35 to 397), while others might be full length bacterial sequences, or even archaeal or eukaryotic sequences.

.

References

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: Inference of RNA alignments. in press.

Walker, J. J. and Pace, N. R. (2007). Phylogenetic composition of rocky mountain endolithic microbial ecosystems. *Appl Environ Microbiol.*, 73:3497–3504.