

coCOA User's Manual

Compositionally Orthogonalized Co-Occurrence Analysis: A program to identify co-occurring motifs in pairs of nucleotide sequences, controlling for G+C heterogeneity.

coCOA executable and compiling instructions

A coCOA PC/Linux executable is provided in src/coCOA. It was compiled on a Linux machine running Fedora.

If you need to compile your own executable, source code and a Makefile for this purpose are provided in the src directory. NOTE: coCOA uses the Rmath standalone library which you must first install before compiling. You then will need to change the -I and -L flags in the Makefile to reflect the path to Rmath.h and libRmath.so, respectively, if they are not in default locations. If they are in default locations you should remove those flags as they may confuse your compiler. To compile, just cd to the src directory and type 'make'. The executable will appear as src/coCOA. Any warnings about the roundl() function can be safely ignored.

Sequence files

The six sequence sets described in the manuscript can be found in the "sequences" directory:

CI.start-end.fa	constitutive intron start/end (human)
SE.flanks.fa	5' end of intron upstream of and 3' end of intron downstream of skipped exons (human)
CE.flanks.fa	5' end of intron upstream of and 3' end of intron downstream of constitutive exons (human)
5ss-p3ss.fa	sequence pairs associated with authentic 5' splice sites and decoy (pseudo) 3' splice site (human)
p5ss-3ss.fa	sequence pairs associated with decoy (pseudo) 5' splice sites and authentic 3' splice sites (human)
mouse.CI.start-end.fa	constitutive intron start/end (mouse)
Promoter-Regions.fa	-420..-201 and -200..+20 relative to the transcription start site of human genes

Running coCOA

`src/coCOA` gives a short usage message. The arguments are as follows:

<code>fasta_fn</code>	Path to sequence filename
<code>nts0</code>	Nucleotides to take from first sequence of each pair. 80 in all examples above, except 220 for <code>Promoter-Regions.fa</code> , but can be smaller to examine only initial segments of sequences, or negative to examine only terminal segments.
<code>nts1</code>	Same as <code>nts0</code> , but for second sequence of each pair.
<code>k</code>	Length of oligomer to examine
<code>nseq</code>	Total number of sequences in file. Note: that is twice the number of sequence pairs. An easy way to calculate this is with <code>grep -c '^>' sequences.fa</code>
<code>outbase</code>	A prefix for the different files that coCOA outputs.

You can ignore warning messages such as the following:

Error: couldn't look up alphabet 'N' (\116)

Error: Couldn't convert kmer [CCGNTAG] to a code

These merely indicate that one of your sequences has an 'N' in it but do not affect the results.

Example:

```
src/coCOA sequences/SE.flanks.fa 80 80 4 25556 example/SE.flanks
```

The output from this command can be found in the example directory.

Output files

You can follow the progress of your process by examining the contents of `outbase.log`, which is updated as the program runs. The limiting factor for runtime is typically the value of `k`, and not the number of sequences. Typically `k=4` will run in a few seconds, `k=5` in a minute or two, and `k=6` in about 30 minutes, regardless of the number of sequences. `outbase.log` will also serve as a record of the run for future reference.

coCOA will create the following files. They are all in the tab-separated-values format, with the first line being tab-separated column names, and the rest of the lines being tab-separated values of those fields. Shown are the filename, a short description of its contents, and a short description of each of the fields in the file.

`outbase.GC_sq0.tsv`

Number of first sequences in each GC bin

GC	GC bin
n	Number of sequences

`outbase.GC_sq1.tsv`

Number of second sequences in each GC bin

GC	GC bin
n	Number of sequences

`outbase.GC_pair.tsv`

Number of sequence pairs in each co-GC bin

GC_sq0	GC bin for first sequence
GC_sq1	GC bin for second sequence
n	Number of sequence pairs

`outbase.kmer_sq0.tsv`

Kmer counts in first sequence set

Kmer	kmer
n	Number of first sequences containing that kmer

`outbase.kmer_sq1.tsv`

Kmer counts in second sequence set

Kmer	kmer
n	Number of second sequences containing that kmer

`outbase.kmer_by_GC_sq0.tsv`

Marginal kmer frequencies in first sequence set

Kmer	kmer (denoted x in the paper)
GC	GC bin of first sequence (b_1)
n	Number of first sequences in this GC-bin containing that kmer ($n_1^{b_1}(x)$)
f_GC	Marginal kmer frequency for that kmer in that GC-bin ($f_1^{b_1}(x)$)

`outbase.kmer_by_GC_sq1.tsv`

Marginal kmer frequencies in second sequence set

Kmer	kmer (denoted y in the paper)
GC	GC bin of first sequence (b_2)
n	Number of first sequences in this GC-bin containing that kmer ($n_2^{b_2}(x)$)
f_GC	Marginal kmer frequency for that kmer in that GC-bin ($f_2^{b_2}(x)$)

outbase.kmer_pairs.tsv

coCOA analysis for all kmer pairs

KmerSq0	kmer for first sequence (denoted x in the paper)
KmerSq1	kmer for second sequence (y)
n	Number of sequence pairs containing x in first sequence and y in the second sequence
muSCOA	Expected number of co-occurrences under null hypothesis of independence between first and second sequence sets (SCOA=simple co-occurrence analysis).
mu	Expected number of co-occurrences under coCOA.
P	Significance level for testing $n > \mu$.

outbase.kmer_pairs.sig.tsv

coCOA analysis for significant kmer pairs

Same fields as outbase.kmer_pairs.tsv, but only contains significant ($P < 4^{-2k}$) co-occurrences. This file forms the basis of subsequent analysis in the paper.

Citation

If you use coCOA in published research please include the following citation:

Friedman BA, Stadler MB, Shomron N, Ding Y, Burge CB, Ab Initio Identification of Functionally Interacting Pairs of cis-Regulatory Elements, 2008.

Questions?

Direct any questions about installing or running coCOA to

Brad Friedman
friedm@mcb.harvard.edu
617-495-5038