Documentation of CRAC software

Nicolas PHILIPPE and Mikaêl SALSON

April 4, 2014

Contents

1 Introduction

As a lot of software, there are many optional parameters in CRAC but only four of them are mandatory (`-k <klength>`, `-i <index>`, `-r <reads1 [reads2]>`, `-o <output_file>`). The output SAM can be generated on the STDOUT by using `-o -` and you can use `--sam` or `--all` instead `-o` argument. This document is intended to guide users of CRAC to choose the more appropriate parameters according to their needs.

2 Selection of a good k-mer length

In CRAC, the value of `-k <int>` is very important for the algorithm. You must not underestimate it, otherwise the results will be of no utility.

For a given genome length and its nucleic composition, a certain sequence length is sufficient to match in average to a unique genomic position with high probability. This length, denoted k, can be computed and optimized. Thus, in a read any k-mer (a k-long substring) can be used as a witness of the possible read matching locations in the genome. In other words, it is possible to use a k-mer in a longer read to find its location without ambiguity on the reference genome.

A k-mer may still have a random match on the reference genome. However, in average over all k-mers, the probability of getting a false location is $< 10^{-4}$, with k = 22, for the Human genome (Philippe et.al, NAR, 2009).

2.1 Suggestion of a k-mer length from different reference genomes

Genome	k
Human	22
Mus musculus	22
Drosophila	20

3 Parameters

3.1 Help options

- `-h`, `--help` to print the principal help page of CRAC.
- `-f`, `--full-help` to print the complete help page of CRAC.
- `-v` to print version of CRAC.

3.2 Mandatory options

- `-i <index\_file>` is the name of the index previously built with the crac-index binary file. Note that crac-index construct the structure `<index\_file.ssa>` with its configuration <index$\backslash_{\text{file}}$.conf> so only the prefix `<index\_file>` must be specified (without extension) to consider the structure and the configuration files both in CRAC.
- `-r <reads\_file1> [reads\_file2]` is the source(s) of the FASTA or FASTQ file(s) containing the reads. Note that the number of files depends if single or paired-reads. The input file may also be compressed using gzip.
- `-k <int>` is the length of the k-mer to be used to map the reads on the reference `<index\_file>`. Note that the condition (k < m) is necessary and reads (or both paired reads) are ignored if m < k. It must be chosen to ensure (as much as possible) that a k-mer has a very high probability to occur a single time on the genome.
- `-o <output\_file>` (or `--sam <output\_file>`) is the output file in SAM format or print on STDOUT with "-o -" argument. See the Documentation of SAM format in CRAC for more details.

3.3 Optional parameters

3.3.1 protocol

- `--stranded` must be specified if reads are produced by a stranded protocol of RNA-Seq (not stranded by default)

3.3.2 efficiency

- `--reads-length <int>` (or `-m <int>`) must be specified for reads of fixed length. If the read length is fixed, we deeply recommend you to specify the read length, by using the `-m` parameter. CRAC will therefore be much faster. `--reads-length <int>` is specified for variable or longer reads, reads shorter are ignored and reads longer are trimmed.

- `--treat-multiple <none>` permits to display alignments with multiple perfectly hits (>max-duplication) rather than a single alignment in the SAM file. Be careful, multiple reads alignments which contain biological events like a SNV, an indel, a splicing or a chimera are not displayed. Indeed, CRAC provides only one alignment by duplicated event (the best one) and the `XD` flag is up. Note that the duplicated events can disallow, specifying the `--no-ambiguity` argument.

- `--nb-threads <int>` is the number of threads to run crac, computational time is almost divided by the number of threads.

- `--max-locs <int>` corresponds to the max number of occurrences retrieved in the index for a given k-mer: smaller is faster, but with a small value, you may miss some locations that would help CRAC detecting the right cause.

3.3.3 accuracy

- `--no-ambiguity <none>` discard biological events (splice, svn, indel, chimera) which have several alignments on the reference index. Indeed, if crac has identified a biological cause in the read that can match in different places of the genome we classify this cause as a biological undetermined event.

3.4 Optional output arguments

Because SAM format is read-centered and, in CRAC's philosophy, we do not see a read as a whole. Rather, it considers portions of the read. Hence, we propose homemade formats of CRAC to classify all breaks for each read. In other words, a same read can be classified several times in different files. See the Documentation of Homemade output format in CRAC for more details.

3.4.1 gz

- `--gz <none>` all output files specified after this argument are gzipped (included for the sam file if `-o/--sam` argument is specified after).

3.4.2 summary

- `--summary <output\_file>` save some statistics about mapping and classification.

3.4.3 all

- `--all <base\_filename>` set output base filename for all causes following. Note that only a `base\_filename` must be specified. Then, the appropriate file extension is added for each cause (SNP, chimera, splice, ...).

3.4.4 normal

- `--normal <output\_file>` save reads that do not contain any break
- `--almost-normal <output\_file>` save reads that do not contain any break but with a variable support

3.4.5 mapping

- `--single <output\_file>` save reads which are located in this way: at least `--min-percent-single-loc <float>` of k-mers are once located on the reference index.
- `--duplicate <output\_file>` save reads which are located in this way: at least `--min-percent-duplication-loc <float>` of k-mers are a few times on the reference index (ie. between `--min-duplication <int>` and `--max-duplication <int>` of locations).
- `--multiple <output\_file>` save reads which are located in this way: at least `--min-percent-multiple-loc <float>` of k-mers are a many times on the reference index (ie. more than `--max-duplication <int>` of locations).
- `--none <output\_file>` save reads which are not located on the reference index.

3.4.6 biological causes

- `--snv <output\_file>` save reads that contain at least a snv.
- `--indel <output\_file>` save reads that contain at least a biological indel.
- `--splice <output\_file>` save reads that contain at least a splicing junction.
- `--weak-splice <output\_file>` save reads that contain at least a low coverage splicing junction.
- `--chimera <output\_file>` save reads that contain at least a chimera junction (junction on different chromosomes, strands or genes).
- `--paired-end-chimera <output\_file>` save paired-end reads that contains a chimera in the non-sequenced part of the original fragment.

- `--biological <output\_file>` save reads that contain a biological cause but for which there is not enough information to be more specific. Note that the biological cause is described for each read.

3.4.7 Sequence errors

- `--errors <output\_file>` save reads that contain at least a sequence error.

3.4.8 Repetition

- `--repeat <output\_file>` save reads that contain a repeated sequence: at least `--min-percent-repetition-loc <float>` percent of consecutive k-mers of a given read are located at least `--min-repetition <int>` occurrences on the reference index. Note that CRAC allows only once repetition by read (the longest).

3.4.9 Other causes

- `--undetermined <output\_file>` save reads that contain an undetermined error: some k-mers are not located on the genome, but the reason for that could not be determined. Note that the error is described for each read.

- `--nothing <output\_file>` save reads that are unclassified.

3.5 Optional process for specific research

- `--deep-snp <none>` must be specified to increase sensitivity to find SNVs at the cost of more computations (only substitution, no indels YET). That process searches for SNV in border cases reads. Those reads would otherwise be classified in bioundetermined.

- `--stringent-chimera <none>` must be specified to increase accuracy to find chimera junctions in exchange of sensitivity and computational times.

3.6 Optional process launcher (once must be selected)

- `--emt <none>` launch an exact matching processing of reads on the index. Either the argument `-k` is equal to 0 which means that the entire read is perfectly mapped on the genome or only a factor of length `k` per read is mapped (the first one with a location) and the rest is sofclipped. With this process, reads are not indexed and it provides a low memory consumption. Note this kind of method is very useful for DGE reads mapping.

- `--server <none>` launch a server to query a given read more precisely. That process is useful for debugging. Note that the output arguments will not be taken into account. Give an `--input-name-server <string>` to set the input fifo name (classify.fifo by default) and give an `--output-name-server`

<string> to set the output fifo name (classify.out.fifo by default). The server can then be used through a client crac-client.

3.7 Additional settings for users

3.7.1 sam output file

- --detailed-sam <none> more information is added in SAM output file. See the Documentation of SAM format for more details.

3.7.2 mapping classification

- --min-percent-single-loc <float> is, to consider a given read as uniquely mapped, the minimum proportion of k-mers that are uniquely mapped on the index (0.15 by default).
- --min-duplication <int> is the minimum number of location to consider a duplicated k-mer (2 by default).
- --max-duplication <int> is the maximum number of location to consider a duplicated k-mer (9 by default).
- --min-percent-duplication-loc <float> is, to consider a given read as duplicated, the minimum proportion of k-mers that are duplicated on the index (0.15 by default).
- --min-percent-multiple-loc <float> is, to consider a given read as "multiple", the minimum proportion of k-mers that are multiple mapped on the index (0.50 by default).
- --min-repetition <int> is the minimum number of locations to consider a repeated k-mer (20 by default).
- --max-percent-repetition-loc <float> is, for a given read, the minimum proportion of k-mers that are repeated on the index to consider a repetition (0.20 by default).

3.7.3 biological causes

- --max-splice-length <int> is the threshold to consider a splice, ie. a splice is reported if the junction length is below max-splice-length <int>, a chimera is considered otherwise (distance by default is 300Kb).
- --max-bio-indel <int> is the threshold to consider a biological indel, ie. an indel is reported if the gap length is below max-bio-indel, a splice is considered otherwise (distance by default is 15).
- --max-bases-retrieved <int> is the number of nucleotides to display in $\text{output}_{\text{file}}$ in case of insertion (15 by default).

3.7.4 errors and undetermined

- `--min-support-no-cover <float>` is the minimum coverage to be able to report a biological cause. Note that if a single read contains a given substitution, it is difficult (if not impossible) to distinguish a sequence error and a biological cause (1.30 by default).

3.8 Additional settings for advanced users

3.8.1 break verification and fusion (merging mirage breaks)

- `--min-break-length <float>` is the minimal break length (as the percentage of k, the k-mer length) so that a cause can be reported. Theoretically, for a given cause, the break length is always >= (k-mer_length - 1). Otherwise, the break may be merged with a close enough break, or the break will be considered as undetermined. (0.5 by default)

- `--max-bases-randomly-matched <int>` A k-mer overlapping an exon-exon junction, for example, may still match on the genome if the overlap is at the end of the read (without loss of generality). This is due to the fact that the nucleotides starting the second exon may be the same as the nucleotides starting the intron. Theoretically, there is a 0.25 probability that we have the same nucleotide at the first position of the intron and the exon. This option specifies how many nucleotides may be matched randomly at most.

- `--max-extension-length <int>` is the maximum number of k-mers extended at each side of a read break. In fact, for a given break, k-mers with false locations can generate false biological causes, so the consistency is checked for each side of the break to discard false k-mers and readjust the good boundaries of the break (10 by default).

3.8.2 threading

- `--nb-tags-info-stored <int>` is a buffer to store information for each thread during the computing phase (1000 by default). This value must be increased if threads work below their real capabilities. With `--nb-threads 15`, CPU usage must be about 1400%.

3.8.3 deep snv search option

- `--nb-nucleotides-snv-comparison <int>` is the minimum k-mer length tolerated for the deep SNVs search (8 by default).

4 Chimera detection

CRAC implements advanced algorithms for detecting chimeras (any types of aberrant transcripts) using RNA-Seq. CRAC can perform two levels of chimera detection:

- chimera junctions can be detected using single read sequence;
- other chimeras can be observed when genome locations from the two read pairs are not consistent.

4.1 Chimera junction

Chimera junctions are detected using the mapping locations of a single read. There is a high level of control in our algorithm to distinguish a chimera junction from a sequence error. An even higher level of control can be performed (at the cost of more computations) using `--stringent-chimera` parameter. A second level of control is also activated when CRAC uses paired-end reads. Indeed, while a chimera junction is found in one read, a verification is done with the second read to look for a mapping concordance and validate or discard the chimera.

4.2 Chimera without junction

Other chimeras can only be detected within paired-end reads. When both reads of a pair are well located (ie. no ambiguity possible), a comparison can be done to compare their locations on the reference. Therefore we can detect a chimera between the two reads if the reads are not colinear.

Beware these chimeras are output in the SAM file in `XP` optional field available (see the SAM documentation of CRAC for more details). They are also output in a homemade file dedicated to that specific purpose with `--paired-end-chimera`.

5 Examples

5.1 Basic example for reads >= 75 bp

```
crac -i humanIndex -k 22 -r reads\_1.fastq -o output.sam --nb-threads
10
```

In that example CRAC is launched on the human genome indexed in humanIndex, with 22-mers on the single reads stored in reads_1.fastq. The output is written in the output.sam file and the program is launched in parallel on 10 threads.

5.2 With more details

```
crac -i humanIndex -k 22 -r reads\_1.fastq -o output.sam --nb-threads
10 --detailed-sam
```

Same example but with `--detailed-sam` to save more information for each read in additional columns of output.sam file.

5.3 With paired reads

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq -o
output.sam --nb-threads 10
```

Same example but with paired reads stored respectively in reads_1.fastq and reads_2.fastq instead of single reads.

5.4 With a strand specific RNA-Seq protocol

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq -o
output.sam --nb-threads 10 --stranded
```

With the `--stranded` if reads provided of a strand specific RNA-Seq protocol.

5.5 With fixed reads length

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq --o
output.sam --nb-threads 10 --stranded --reads-length 200
```

With `--reads-length 200` to only consider reads with 200bp-long (or are truncated if longer, or ignored if shorter).

5.6 With a min break length chosen

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq --o
output.sam --nb-threads 10 --stranded --reads-length 200 --min-break-length
0.7
```

With `--min-break-length 0.7` to reject break with a reduction more than 70% relative to the theoretical break length.

5.7 With an other species and reads < 75 bp

```
crac -i drosophiliaIndex -k 20 -r reads.fastq -m 50 --all output
--nb-threads 10 --max-extension-length 5
```

In that example CRAC is launched on the Drosophila genome indexed in drosophiliaIndex, with 20-mers on the reads stored in reads.fastq (single reads format). All reads are 50bp-long (or are truncated if longer, or ignored if shorter). The output files is written both in a sam format and in each category file format (output.snv, output.chimera, etc) and the program is launched in parallel on 10 threads. Here, the `max-extension-length` is decreased to avoid a loss en sensitivity on very short reads.

5.8 Now, for specific research of chimeras

5.8.1 with accuracy

```
crac -i humanIndex -k 22 -r reads\_1.fastq -o output.sam --nb-threads
10 --chimera output.chimera --stringent-chimera
```

In that example CRAC is launched on the human genome indexed in humanIndex, with 22-mers on the single reads stored in reads_1.fastq. The output is written

in the output.sam file and the program is launched in parallel on 10 threads. Information for reads that contain chimera is stored in output.chimera. The stringent process `--stringent-chimera` is applied on chimeras to increase the accuracy in exchange of sensitivity and computational times.

5.8.2 with paired-end control

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq -o
output.sam --nb-threads 10 --chimera output.chimera --stringent-chimera
```

This example is the same as the one before, except that we are using paired-end reads. In this case CRAC will perform an extra test to discard chimera junctions that don't match paired-end read mapping.

5.8.3 with research of non-sequenced chimera junctions

```
crac -i humanIndex -k 22 -r reads\_1.fastq reads\_2.fastq -o
output.sam --nb-threads 10 --paired-end-chimera output.chimera
```

This example performs the research of chimera junctions that are not directly sequenced but are highlighted within paired-end reads mapping. Those chimeras are output in the file `output.chimera`.