Gfapy Documentation

Release 1.0.0

Giorgio Gonnella

Mar 20, 2017

CONTENTS:

CHAPTER

ONE

INTRODUCTION

The Graphical Fragment Assembly (GFA) are formats for the representation of sequence graphs, including assembly, variation and splicing graphs. Two versions of GFA have been defined (GFA1 and GFA2) and several sequence analysis programs have been adopting the formats as an interchange format, which allow to easily combine different sequence analysis tools.

This library implements the GFA1 and GFA2 specification described at https://github.com/GFA-spec/GFA-spec/blob/master/GFA-spec.md. It allows to create a Gfa object from a file in the GFA format or from scratch, to enumerate the graph elements (segments, links, containments, paths and header lines), to traverse the graph (by traversing all links outgoing from or incoming to a segment), to search for elements (e.g. which links connect two segments) and to manipulate the graph (e.g. to eliminate a link or a segment or to duplicate a segment distributing the read counts evenly on the copies).

The GFA format can be easily extended by users by defining own custom tags and record types. In Gfapy, it is easy to write extensions modules, which allow to define custom record types and datatypes for the parsing and validation of custom fields. The custom lines can be connected, using references, to each other and to lines of the standard record types.

1.1 Requirements

Gfapy has been written for Python 3 and tested using Python version 3.3. It does not require any additional Python packages or other software.

1.2 Installation

Gfapy is distributed as a Python package and can be installed using the python package manager pip.

The following command installs the current stable version from the Python Packages index:

```
pip install gfapy
```

If you would like to install the current development version from Github, use the following command:

```
pip install -e git+https://github.com/ggonnella/gfapy.git#egg=gfapy
```

1.3 Usage

If you installed gfapy as described above, you can import it in your script using the conventional Python syntax:

```
>>> import gfapy
```

1.4 Documentation

An user manual is available at https://github.com/ggonnella/gfapy/blob/master/manual/gfapy-manual.pdf

1.5 References

The manuscript describing Gfapy has been submitted and is currently under review. This section will be updated, as soon as the publication is available.

CHAPTER

TWO

CHANGELOG

```
== 1.0.0 ==

- initial release
```

CHAPTER

THREE

THE GFA CLASS

The content of a GFA file is represented in Gfapy by an instance of the class `Gfa`. In most cases, the Gfa instance will be constructed from the data contained in a GFA file, using the method `Gfa.from_file()`.

Alternatively, it is possible to use the construct of the class; it takes an optional positional parameter, the content of a GFA file (as string, or as list of strings, one per line of the GFA file). If no GFA content is provided, the Gfa instance will be empty.

```
>>> gfa = gfapy.Gfa("H\tVN:Z:1.0\nS\tA\t*")
>>> print(len(gfa.lines))
2
>>> gfa = gfapy.Gfa(["H\tVN:Z:1.0", "S\tA\t*", "S\tB\t*"])
>>> print(len(gfa.lines))
3
>>> gfa = gfapy.Gfa()
>>> print(len(gfa.lines))
0
```

The string representation of the Gfa object (which can be obtained using `str()`) is the textual representation in GFA format. Using `Gfa.to_file(filename)` allows writing this representation to a GFA file (the content of the file is overwritten).

```
>>> g1 = gfapy.Gfa()
>>> g1.append("H\tVN:Z:1.0")
<gfapy.gfa.Gfa object at 0x...>
>>> g1.append("S\ta\t*")
<gfapy.gfa.Gfa object at 0x...>
>>> g1.to_file("my.gfa")
>>> g2 = gfapy.Gfa.from_file("my.gfa")
>>> str(g1)
'H\tVN:Z:1.0\nS\ta\t*'
```

All methods for creating a Gfa (constructor and from_file) accept a `vlevel` parameter, the validation level, and can assume the values 0, 1, 2 and 3. An higher value means more validations are performed. The Validations chapter explains the meaning of the different validation levels in detail. The default value is 1.

```
>>> gfapy.Gfa().vlevel
1
>>> gfapy.Gfa(vlevel = 0).vlevel
0
```

A further parameter is `version`. It can be set to `'gfa1'`, `'gfa2'` or left to the default value (`None`). The default is to auto-detect the version of the GFA from the line content. If the version is set manually, any content not compatible to the specified version will trigger an exception. If the version is set automatically, an exception will be raised if two lines are found, with content incompatible to each other (e.g. a GFA1 segment followed by a GFA2 segment).

```
>>> g = gfapy.Gfa(version='gfa2')
>>> g.version
```

```
'gfa2'
>>> g.add_line("S\t1\t*")
Traceback (most recent call last):
...
gfapy.error.VersionError: Version: 1.0 (None)
...
>>> g = gfapy.Gfa()
>>> g.version
>>> g.add_line("S\t1\t*")
<gfapy.gfa.Gfa object at ...>
>>> g.version
'gfa1'
>>> g.add_line("S\t1\t100\t*")
Traceback (most recent call last):
...
gfapy.error.VersionError: Version: 1.0 (None)
...
```

3.1 Collections of lines

The property `lines` of the Gfa object is a list of all the lines in the GFA file (including the header, which is splitted into single-tag lines). The list itself shall not be modified by the user directly (i.e. adding and removing lines is done using a different interface, see below). However the single elements of the list can be edited.

```
>>> for line in gfa.lines: print(line)
```

For most record types, a list of the lines of the record type is available as a read-only property, which is named after the record type, in plural.

```
>>> [str(line) for line in gfa1.segments]
['S\t1\t*','S\t2\t*','S\t3\t*']
>>> [str(line) for line in gfa2.fragments]
[]
```

A particular case are edges; these are in GFA1 links and containments, while in GFA2 there is an unified edge record type, which also allows to represent internal alignments. In Gfapy, the `edges` property retrieves all edges (i.e. all E lines in GFA2, and all L and C lines in GFA1). The `dovetails` property is a list of all edges which represent dovetail overlaps (i.e. all L lines in GFA1 and a subset of the E lines in GFA2). The `containments` property is a list of all edges which represent containments (i.e. all C lines in GFA1 and a subset of the E lines in GFA2).

```
>>> gfa2.edges
[]
>>> gfa2.dovetails
[]
>>> gfa2.containments
[]
```

Paths are retrieved using the `paths` property: this list contains all P lines in GFA1 and all O lines in GFA2. Sets returns the list of all U lines in GFA2 (empty list in GFA1).

```
>>> gfa2.paths
[]
>>> gfa2.sets
[]
```

The header contain metadata in a single or multiple lines. For ease of access to the header information, all its tags are summarized in a single line instance, which is retrieved using the read-only `header` property. The Header chapter of this manual explains more in detail, how to work with the header object.

```
>>> gfa2.header.TS
100
```

All lines which start by the string # are comments; they are handled in the "Comments" chapter and are retrieved using the `comments` property:

```
>>> [str(line) for line in gfa1.comments]
['# this is a comment']
```

Custom lines are lines of GFA2 files which start with a non-standard record type. Gfapy provides basic built-in support for accessing the information in custom lines, and allows to define extensions for own record types for defining more advanced functionality (described in the Supplemental Information to the manuscript presenting gfapy).

```
>>> [str(line) for line in gfa2.custom_records]
['X\tcustom line', 'Y\tcustom line']
>>> gfa2.custom_record_keys)
['X', 'Y']
>>> [str(line) for line in gfa2.custom_records_of_type('X')]
['X\tcustom line']
```

3.2 Line identifiers

Some GFA lines have a mandatory or optional identifier field: segments and paths in GFA1, segments, gaps, edges, paths and sets in GFA2. A line of this type can be retrieved by identifier, using the method `Gfa.line(ID)` using the identifier as argument.

The list of all identifier can be retrieved using the `names` property; for the identifiers of a single line type, a property is available, named after the record type in singular, with the `_names` suffix:

```
>>> str(gfa1.line('1'))
'S\t1\t*'
```

The list of all identifier can be retrieved using the `names` property; for the identifiers of a single line type, a property is available, named after the record type in singular, with the `_names` suffix. Segment names and path names are for both GFA versions, while edge, gap and set names will always be empty lists in GFA1 Gfa instances.

```
>>> g = gfapy.Gfa()
>>> g = g.add_line("S\tA\t*")
>>> g.names
['A']
>>> g.segment_names
['A']
>>> g.path_names
[]
>>> g.edge_names
[]
>>> g.gap_names
[]
>>> g.set_names
[]
```

3.3 Identifiers of external sequences

Fragments contain identifiers which refer to external sequences (not contained in the GFA file). According to the specification, the these identifiers are not part of the same namespace as the identifier of the GFA lines. They can be retrieved using the `external_names` property:

```
>>> g.external_names
[]
```

The method `Gfa.fragments_for_external(external_ID)` retrieves all F lines with a specified external sequence identifier.

3.4 Adding new lines

New lines can be added to a Gfa instance using the `Gfa.add_line(line)` method or its alias `Gfa.append(line)`. The argument can be either a string describing a line with valid GFA syntax, or a `Line` instance. If a string is added, a line instance is created and then added.

3.5 Editing the lines

Accessing the information stored in the fields of a line instance is described in the "Positional fields" and "Tags" chapters.

In Gfapy, a line instance belonging to a Gfa instance is said to be *connected* to the Gfa instance. Direct editing the content of a connected line is only possible, for those fields which do not contain references to other lines. For more information on how to modify the content of the fields of connected line, see the "References" chapter.

3.6 Removing lines

Disconnecting a line from the Gfa instance is done using the `rm(line)` method. The argument can be a line instance or a string (in which case the line is searched using the `line(name)` method, then eliminated). A line instance can also be disconnected using the `disconnect()` method on it. Disconnecting a line may trigger other operations, such as the disconnection of other lines (see the "References" chapter).

3.7 Renaming lines

Lines with an identifier can be renamed. This is done simply by editing the corresponding field (such as `name` or `sid` for a segment). This field is not a reference to another line and can be freely edited also in line instances connected to a Gfa. All references to the line from other lines will still be up to date, as they will refer to the same instance (whose name has been changed) and their string representation will use the new name.

CHAPTER

FOUR

VALIDATION

Different validation levels are available. They represent different compromises between speed and warrant of validity. The validation level can be specified when the `gfapy.Gfa` object is created, using the `vlevel` parameter of the constructor and of the `gfapy.Gfa.from_file()` method. Four levels of validation are defined (0 = no validation, 1 = validation by reading, 2 = validation by reading and writing, 3 = continuous validation). The default validation level value is 1.

4.1 Manual validation

Independently from the validation level choosen, the user can always check the value of a field calling `validate_field(fieldname)` on the line instance. If no exeption is raised, the field content is valid.

To check if the entire content of the line is valid, the user can call `validate` on the line instance. This will check all fields and perform cross-field validations, such as comparing the length of the sequence of a GFA1 segment, to the value of the LN tag (if present).

It is also possible to validate the structure of the GFA, for example to check if there are unresolved references to lines. To do this, use the `validate()` method of the `gfapy.Gfa` class.

4.2 No validations

If the validation is set to 0, Gfapy will try to accept any input and never raise an exception. This is not always possible, and in some cases, an exception will still be raised, if the data is invalid.

4.3 Validation when reading

If the validation level is set to 1 or higher, basic validations will be performed, such as checking the number of positional fields, the presence of duplicated tags, the tag datatype of predefined tags. Additionally, all tags will be validated, either during parsing or on first access. Record-type cross-field validations will also be performed.

In other words, a validation of 1 means that Gfapy guarantees (as good as it can) that the GFA content read from a file is valid, and will raise an exception on accessing the data if not.

The user is supposed to run `validate_field(fieldname)` when changing a field content to something which can be potentially invalid, or `validate()` if potentially cross-field validations could fail.

4.4 Validation when writing

Setting the level to 2 will perform all validations described above, plus validate the fields content when their value is written to string.

In other words, a validation of 2 means that Gfapy guarantee (as good as it can) that the GFA content read from a file and written to a file is valid and will raise an exception on accessing the data or writing to file if not.

4.5 Continuous validation

If the validation level is set to 3, all validations for lower levels described above are run, plus a validation of fields contents each time a setter method is used.

A validation of 3 means that Gfapy guarantees (as good as it can) that the GFA content is always valid.

CHAPTER

FIVE

POSITIONAL FIELDS

Most lines in GFA have positional fields (Headers are an exception). During parsing, if a line is encountered, which has too less or too many positional fields, an exception will be thrown. The correct number of positional fields is record type-specific.

Positional fields are recognized by its position in the line. Each positional field has an implicit field name and datatype associated with it.

5.1 Field names

The field names are derived from the specification. Lower case versions of the field names are used and spaces are subsituted with underscores. In some cases, the field names were changed, as they represent keywords in common programming languages (`from`, `send`).

The following tables shows the field names used in Gfapy, for each kind of line. Headers have no positional fields. Comments and custom lines follow particular rules, see the respective chapters.

5.1.1 GFA1 field names

Record Type	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6
Segment	`name`	`sequence`				
Link	`from_segment`	`from_orient`	`to_segment`	`to_orient`	`overlap`	
Containment	`from_segment`	`from_orient`	`to_segment`	`to_orient`	`pos`	`overlap`
Path	`path_name`	`segment_names`	`overlaps`			

5.1.2 GFA2 field names

Record Type	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8
Segment	`sid`	`slen`	`sequence`					
Edge	`eid`	`sid1`	`sid2`	`beg1`	`end1`	`beg2`	`end2`	`alignment`
Fragment	`sid`	`external`	`s_beg`	`s_end`	`f_beg`	`f_end`	`alignment`	
Gap	`gid`	`sid1`	`d1`	`d2`	`sid2`	`disp`	`var`	
Set	`pid`	`items`						
Path	`pid`	`items`						

5.2 Datatypes

The datatype of each positional field is described in the specification and cannot be changed (differently from tags). Here is a short description of the Python classes used to represent data for different datatypes. For some

complex cases, more details are found in the following chapters.

5.2.1 Placeholders

The positional fields in GFA can never be empty. However, there are some fields with optional values. If a value is not specified, a placeholder character is used instead (*). Such undefined values are represented in Gfapy by the `gfapy.Placeholder` class, which is described more in detail in the Placeholders chapter.

5.2.2 Arrays

The `items` field in unordered and ordered groups and the `segment_names` and `overlaps` fields in paths are lists of objects and are represented by list instances.

```
type(set.items) # => "list"
type(gfa2_path.items) # => "list"
type(gfa1_path.segment_names) # => "list"
type(gfa1_path.overlaps) # => "list"
```

5.2.3 Orientations

Orientations are represented by strings. The `gfapy.invert()` method applied to an orientation string returns the other orientation.

```
>>> gfapy.invert("+")
'-'
>>> gfapy.invert("-")
'+'
```

5.2.4 Identifiers

The identifier of the line itself (available for S, P, E, G, U, O lines) can always be accessed in Gfapy using the `name` alias and is represented in Gfapy by a string. If it is optional (E, G, U, O lines) and not specified, it is represented by a Placeholder instance. The fragment identifier is also a string.

Identifiers which refer to other lines are also present in some line types (L, C, E, G, U, O, F). These are never placeholders and in stand-alone lines are represented by strings. In connected lines they are references to the Line instances to which they refer to (see the References chapter).

5.2.5 Oriented identifiers

Oriented identifiers (e.g. `segment_names` in GFA1 paths) are represented by elements of the class `gfapy.OrientedLine`. The `segment` method of the oriented segments returns the segment identifier (or segment reference in connected path lines) and the `orient` method returns the orientation string. The `name` method returns the string of the segment, even if this is a reference to a segment. A new oriented line can be created using the `OL[line, orientation]` method.

Calling `invert` returns an oriented segment, with inverted orientation. To set the two attributes the methods `segment=` and `orient=` are available.

Examples:

```
>>> p = gfapy.Line.from_string("P\tP1\ta+,b-\t*")
>>> p.segment_names
[gfapy.OrientedLine('a','+'), gfapy.OrientedLine('b','-')]
>>> sn0 = p.segment_names[0]
>>> sn0.line
```

```
'a'
>>> sn0.name
'a'
>>> sn0.orient
'+'
>>> sn0.invert()
>>> sn0
gfapy.OrientedLine('a','-')
>>> sn0.orient
'-'
>>> sn0.line = gfapy.Line.from_string('S\tX\t*')
>>> str(sn0)
'X-'
>>> sn0.name
'X'
>>> sn0 = gfapy.OrientedLine(gfapy.Line.from_string('S\tY\t*'), '+')
```

5.2.6 Sequences

Sequences (S field sequence) are represented by strings in Gfapy. Depending on the GFA version, the alphabet definition is more or less restrictive. The definitions are correctly applied by the validation methods.

The method `rc()` is provided to compute the reverse complement of a nucleotidic sequence. The extended IUPAC alphabet is understood by the method. Applied to non nucleotidic sequences, the results will be meaningless:

```
>>> from gfapy.sequence import rc
>>> rc("gcat")
'atgc'
>>> rc("*")
'*'
>>> rc("yatc")
'gatr'
>>> rc("gCat")
'atGc'
>>> rc("cag", rna=True)
'cug'
```

5.2.7 Integers and positions

The C lines `pos` field and the G lines `disp` and `var` fields are represented by integers. The `var` field is optional, and thus can be also a placeholder. Positions are 0-based coordinates.

The position fields of GFA2 E lines (`beg1, beg2, end1, end2`) and F lines (`s_beg, s_end, f_beg, f_end`) contain a dollar string as suffix if the position is equal to the segment length. For more information, see the Positions chapter.

5.2.8 Alignments

Alignments are always optional, ie they can be placeholders. If they are specified they are CIGAR alignments or, only in GFA2, trace alignments. For more details, see the Alignments chapter.

5.2.9 GFA1 datatypes

Datatype	Record Type	Fields
Identifier	Segment	`name`
	Path	`path_name`
	Link	`from_segment, to_segment`
	Containment	`from_segment, to_segment`
[OrientedIdentifier]	Path	`segment_names`
Orientation	Link	`from_orient, to_orient`
	Containment	`from_orient, to_orient`
Sequence	Segment	`sequence`
Alignment	Link	`overlap`
	Containment	`overlap`
[Alignment]	Path	`overlaps`
Position	Containment	`pos`

5.2.10 GFA2 datatypes

Datatype	Record Type	Fields
Itentifier	Segment	`sid`
	Fragment	`sid`
OrientedIdentifier	Edge	`sid1, sid2`
	Gap	`sid1, sid2`
	Fragment	`external`
OptionalIdentifier	Edge	`eid`
	Gap	`gid`
	U Group	`oid`
	O Group	`uid`
[Identifier]	U Group	`items`
[OrientedIdentifier]	O Group	`items`
Sequence	Segment	`sequence`
Alignment	Edge	`alignment`
	Fragment	`alignment`
Position	Edge	`beg1, end1, beg2, end2`
	Fragment	`s_beg, s_end, f_beg, f_end`
Integer	Gap	`disp, var`

5.3 Reading and writing positional fields

The `positional_fieldnames` method returns the list of the names (as strings) of the positional fields of a line. The positional fields can be read using a method on the Gfapy line object, which is called as the field name. Setting the value is done with an equal sign version of the field name method (e.g. segment.slen = 120). In alternative, the `set(fieldname, value)` and `get(fieldname)` methods can also be used.

```
>>> s_gfa1 = gfapy.Line.from_string("S\t1\t*")
>>> s_gfa1.positional_fieldnames
['name', 'sequence']
>>> s_gfa1.name
'1'
>>> s_gfa1.get("name")
'1'
>>> s_gfa1.name = "segment2"
>>> s_gfa1.name
'segment2'
>>> s_gfa1.set('name',"3")
```

```
>>> s_gfa1.name
'3'
```

When a field is read, the value is converted into an appropriate object. The string representation of a field can be read using the `field_to_s(fieldname)` method.

```
link.from_segment # => gfapy.line.segment.GFA1("S\ts1\t*")
link.field_to_s(from_segment) # => ("s1")
```

When setting a non-string field, the user can specify the value of a tag either as a Python non-string object, or as the string representation of the value.

```
c.pos = 1
c.pos = "1"
c.pos # => 1
c.field_to_s("pos") # => "1"
```

Note that setting the value of reference and backreferences-related fields is generally not allowed, when a line instance is connected to a Gfapy object (see the References chapter).

```
s = gfa.Line.from_string("L\ts1\t+\ts2\t-\t*")
s.from_segment = "s3"
gfa.add_line(s)
s.from_segment = "s4" # raises an exception
```

5.4 Validation

The content of all positional fields must be a correctly formatted string according to the rules given in the GFA specifications (or a Python object whose string representation is a correctly formatted string).

Depending on the validation level, more or less checks are done automatically (see the Validation chapter). Not regarding which validation level is selected, the user can trigger a manual validation using the `validate_field(fieldname)` method for a single field, or using `validate`, which does a full validation on the whole line, including all positional fields.

```
line.validate_field("xx")
line.validate()
```

5.5 Aliases

For some fields, aliases are defined, which can be used in all contexts where the original field name is used (i.e. as parameter of a method, and the same setter and getter methods defined for the original field name are also defined for each alias, see below).

```
gfa1_path.name == gfa1_path.path_name # True
edge.eid == edge.name # True
segment.sid == segment.name # True
containment.from_segment == containment.container # True

s = gfapy.Line.from_string("S\t1\t*")
s.sid # => "1"
s.name = "a"
s.sid # => "a"
```

5.5.1 Name

Different record types have an identifier field: segments (name in GFA1, sid in GFA2), paths (path_name), edge (eid), fragment (sid), gap (gid), groups (pid).

All these fields are aliased to `name`. This allows the user for example to set the identifier of a line using the `name=(value)` method using the same syntax for different record types (segments, edges, paths, fragments, gaps and groups).

5.5.2 Version-specific field names

For segments the GFA1 name and the GFA2 sid are equivalent fields. For this reason an alias `sid` is defined for GFA1 segments and `name` for GFA2 segments.

5.5.3 Crypical field names

The definition of from and to for containments is somewhat cryptical. Therefore following aliases have been defined for containments: container[_orient] for from[_|segment|orient]; contained[_orient] for to[_segment|orient].

CHAPTER

SIX

PLACEHOLDERS

Some positional fields may contain an undefined value S: `sequence`; L/C: `overlap`; P: `overlaps`; E: `eid`, `alignment`; F: `alignment`; G: `gid`, `var`; U/O: `pid`. In GFA this value is represented by a `*`.

In Gfapy the class Placeholder represent the undefined value.

6.1 Distinguishing placeholders

The method ‘‘‘gfapy.is_placeholder()‘‘‘‘ checks if a value is or would be represented by a placeholder in GFA (such as an empty array, or a string containing “*”).

```
gfapy.is_placeholder("*") # => True
gfapy.is_placeholder("**") # => False
gfapy.is_placeholder([]) # => True
gfapy.is_placeholder(gfapy.Placeholder()) # => True
```

Note that, as a placeholder is False in boolean context, just a `if not placeholder` will also work, if placeholder is a gfa.Placeholder() but not if it is a string representation.

6.2 Compatibility methods

Some methods are defined for placeholders, which allow them to respond to the same methods as defined values. This allows to write generic code.

```
placeholder.validate() # does nothing
len(placeholder) # => 0
placeholder[1] # => gfapy.Placeholder()
placeholder + anything # => gfapy.Placeholder()
```

CHAPTER

SEVEN

POSITION FIELDS

The only position field in GFA1 is the `pos` field in the C lines. This represents the starting position of the contained segment in the container segment and is 0-based.

Some fields in GFA2 E lines (`beg1`, `beg2`, `end1`, `end2`) and F lines (`s_beg`, `s_end`, `f_beg`, `f_end`) are positions. According to the specification, they are 0-based and represent virtual ticks before and after each string in the sequence. Thus ranges are represented similarly to the Python range conventions: e.g. a 1-character prefix of a sequence will have begin 0 and end 1.

7.1 GFA2 last position string

The GFA2 positions must contain an additional string (`$`) appended to the integer, if (and only if) they are the last position in the segment sequence. These particular positions are represented in Gfapy as instances of the class `gfapy.LastPos`.

To create a lastpos instance, the constructor can be used with an integer, or the string representation (which must end with the dollar sign, otherwise an integer is returned):

```
str(gfapy.LastPos(12))    # => "12$"
gfapy.LastPos("12")       # => 12
str(gfapy.LastPos("12")) # => "12"
gfapy.LastPos("12$")      # => gfapy.LastPos(12)
str(gfapy.LastPos("12$")) # => "12$"
```

Subtracting an integer from a lastpos returns a lastpos if 0 subtracted, an integer otherwise. This allows to do some arithmetic on positions without making them invalid.

```
gfapy.LastPos(12) - 0 # => gfapy.LastPos(12)
gfapy.LastPos(12) - 1 # => 11
```

The functions `gfapy.islastpos` and ‘‘isfirstpos‘‘‘ allow to determine if a position value is 0 (first), or the last position, using the same syntax for lastpos and integer instances.

```
gfapy.isfirst(0)  # True
gfapy.islast(0)   # False
gfapy.isfirst(12) # False
gfapy.islast(12)  # False
gfapy.islast(gfapy.LastPos("12"))  # False
gfapy.islast(gfapy.LastPos("12$")) # True
```

CHAPTER

EIGHT

ALIGNMENTS

Some GFA1 (L/C overlap, P overlaps) and GFA2 (E/F alignment) fields contain alignments or lists of alignments. The alignment can be left unspecified and a placeholder symbol `*` used instead. In GFA1 the alignments can be given as CIGAR strings, in GFA2 also as Dazzler traces.

Gfapy uses three different classes for representing the content of alignment fields: `CIGAR`, `Trace` and `AlignmentPlaceholder`.

8.1 Creating an alignment

An alignment instance is usually created from its GFA string representation or from a list by using the `gfapy.Alignment()` constructor.

```
>>> from gfapy import Alignment
>>> Alignment("*")
gfapy.AlignmentPlaceholder()
>>> Alignment("10,10,10")
gfapy.Trace([10,10,10])
>>> Alignment([10,10,10])
gfapy.Trace([10,10,10])
>>> Alignment("30M2I")
gfapy.CIGAR([gfapy.CIGAR.Operation(30,'M'), gfapy.CIGAR.Operation(2,'I')])
```

If the argument is an alignment object it will be returned, so that is always safe to call the method on a variable which can contain a string or an alignment instance:

```
>>> Alignment(Alignment("*"))
gfapy.AlignmentPlaceholder()
>>> Alignment(Alignment("10,10"))
gfapy.Trace([10,10])
```

8.2 Recognizing undefined alignments

The `gfapy.is_placeholder()` method allows to test if an alignment field contains an undefined value (placeholder) instead of a defined value (CIGAR string, trace). The method accepts as argument either an alignment object or a string or list representation.

```
>>> from gfapy import is_placeholder, Alignment
>>> is_placeholder(Alignment("30M"))
False
>>> is_placeholder(Alignment("10,10"))
False
>>> is_placeholder(Alignment("*"))
True
>>> is_placeholder("*")
```

```
True
>>> is_placeholder("30M")
False
>>> is_placeholder("10,10")
False
>>> is_placeholder([])
True
>>> is_placeholder([10,10])
False
```

Note that, as a placeholder is `False` in boolean context, just a `if not aligment` will also work, if alignment is an alignment object. But this of course, does not work, if it is a string representation. Therefore it is better to use the `gfapy.is_placeholder()` method, which works in both cases.

```
>>> if not Alignment("*"): print('no alignment')
no alignment
>>> if is_placeholder(Alignment("*")): print('no alignment')
no alignment
>>> if "*": print('not a placeholder...?')
not a placeholder...?
>>> if is_placeholder("*"): print('really? it is a placeholder!')
really? it is a placeholder!
```

8.3 Reading and editing CIGARs

CIGARs are represented by specialized lists, instances of the class `CIGAR`, whose elements are CIGAR operations CIGAR operations are represented by instance of the class `Operation`, and provide the properties `length` (lenght of the operation, an integer) and `code` (one-letter string which specifies the type of operation). Note that not all operations allowed in SAM files (for which CIGAR strings were first defined) are also meaningful in GFA and thus GFA2 only allows the operations `M`, `I`, `D` and `P`.

```
>>> cigar = gfapy.Alignment("30M")
>>> isinstance(cigar, list)
True
>>> operation = cigar[0]
>>> type(operation)
<class 'gfapy.alignment.cigar.CIGAR.Operation'>
>>> operation.code
'M'
>>> operation.code = 'D'
>>> operation.length
30
>>> len(operation)
30
>>> str(operation)
'30D'
```

As a CIGAR instance is a list, list methods apply to it. If the array is emptied, its string representation will be the placeholder symbol `*`.

```
>>> cigar = gfapy.Alignment("1I20M2D")
>>> cigar[0].code = "M"
>>> cigar.pop(1)
gfapy.CIGAR.Operation(20,'M')
>>> str(cigar)
'1M2D'
>>> cigar[:] = []
>>> str(cigar)
'*'
```

The validate `CIGAR.validate()` function checks if a CIGAR instance is valid. A version can be provided, as the CIGAR validation is version specific (as GFA2 forbids some CIGAR operations).

```
>>> cigar = gfapy.Alignment("30M10D20M5I10M")
>>> cigar.validate()
>>> cigar[1].code = "L"
>>> cigar.validate()
Traceback (most recent call last):
  ...
gfapy.error.ValueError:
>>> cigar = gfapy.Alignment("30M10D20M5I10M")
>>> cigar[1].code = "X"
>>> cigar.validate(version="gfa1")
>>> cigar.validate(version="gfa2")
Traceback (most recent call last):
  ...
gfapy.error.ValueError:
```

8.4 Reading and editing traces

Traces are arrays of non-negative integers. The values are interpreted using a trace spacing value. If traces are used, a trace spacing value must be defined in a TS integer tag, either in the header, or in the single lines which contain traces (which takes precedence over the header global value).

```
>>> print(gfa)
H TS:i:100
E x A+ B- 0 100$ 0 100$ 4,2 TS:i:50
...
>>> gfa.header.TS
100
>>> gfa.line("x").TS
50
```

8.5 Query, reference and complement

CIGARs are asymmetric, i.e.they consider one sequence as reference and another sequence as query.

The `length_on_reference()` and `length_on_query()` methods compute the length of the alignment on the two sequences. These methods are used by the library e.g. to convert GFA1 L lines to GFA2 E lines (which is only possible if CIGARs are provided).

```
>>> cigar = gfapy.Alignment("30M10D20M5I10M")
>>> cigar.length_on_reference()
70
>>> cigar.length_on_query()
65
```

CIGARs are dependent on which sequence is taken as reference and which is taken as query. For each alignment, a complement CIGAR can be computed using the method `complement()`; it is the CIGAR obtained when the two sequences are switched.

```
>>> cigar = gfapy.Alignment("2M1D3M")
>>> str(cigar.complement())
'3M1I2M'
```

The current version of Gfapy does not provide a way to compute the alignment, thus the trace information can be accessed and edited, but not used for this purpose. Because of this there is currently no way in Gfapy to compute a complement trace (trace obtained when the sequences are switched).

```
>>> trace = gfapy.Alignment("1,2,3")
>>> str(trace.complement())
'*'
```

The complement of a placeholder is a placeholder:

```
>>> str(gfapy.Alignment("*").complement())
'*'
```

CHAPTER

NINE

TAGS

Each record in GFA can contain tags. Tags are fields which consist in a tag name, a datatype and data. The format is `NN:T:DATA` where `NN` is a two-letter tag name, `T` is an one-letter datatype string and `DATA` is a string representing the data according to the specified datatype. Tag names must be unique for each line, i.e. each line may only contain a tag once.

```
# Examples of GFA tags of different datatypes:
"aa:i:-12"
"bb:f:1.23"
"cc:Z:this is a string"
"dd:A:X"
"ee:B:c,12,3,2"
"ff:H:122FA0"
'gg:J:["A","B"]'
```

9.1 Custom tags

Some tags are explicitely defined in the specification (these are named *predefined tags* in Gfapy), and the user or an application can define its own custom tags.

Custom tags are user or program specific and may of course collide with the tags used by other users or programs. For this reasons, if you write scripts which employ custom tags, you should always check that the values are of the correct datatype and plausible.

```
if line.get_datatype("xx") != "i":
  raise Exception("I expected the tag xx to contain an integer!")
myvalue = line.xx
if (myvalue > 120) or (myvalue % 2 == 1):
  raise Exception("The value in the xx tag is not an even value <= 120")
# ... do something with myvalue
```

Also it is good practice to allow the user of the script to change the name of the custom tags. For example, Gfapy employs the +or+ custom tag to track the original segment from which a segment in the final graph is derived. All methods which read or write the +or+ tag allow to specify an alternative tag name to use instead of +or+, for the case that this name collides with the custom tag of another program.

```
# E.g. a method which does something with myvalue, usually stored in tag xx
# allows the user to specify an alternative name for the tag
def mymethod(line, mytag="xx"):
  myvalue = line.get(mytag)
  # ...
```

9.2 Tag names in GFA1

According to the GFA1 specification, custom tags are lower case, while predefined tags are upper case (in both cases the second character in the name can be a number). There is a number of predefined tags in the specification, different for each kind of line.

```
"VN:Z:1.0" # VN is upcase => predefined tag
"z5:Z:1.0" # z5 first char is downcase => custom tag

# not forbidden, but not reccomended:
"zZ:Z:1.0" # => mixed case, first char downcase => custom tag
"Zz:Z:1.0" # => mixed case, first char upcase => custom tag
"vn:Z:1.0" # => same name as predefined tag, but downcase => custom tag
```

Besides the tags described in the specification, in GFA1 headers, the TS tag is allowed, in order to simplify the translation of GFA2 files.

9.3 Tag names in GFA2

The GFA2 specification is currently not as strict regarding tags: anyone can use both upper and lower case tags, and no tags are predefined except for VN and TS.

However, Gfapy follows the same conventions as for GFA1: i.e. it allows the tags specified as predefined tags in GFA1 to be used also in GFA2. No other upper case tag is allowed in GFA2.

9.4 Datatypes

The following table summarizes the datatypes available for tags:

Symbol	Datatype	Example	Python class
Z	string	This is a string	str
i	integer	-12	int
f	float	1.2E-5	float
A	char	X	str
J	JSON	[1,{“k1”:1,”k2”:2},”a”]	list/dict
B	numeric array	f,1.2,13E-2,0	gfapy.NumericArray
H	byte array	FFAA01	gfapy.ByteArray

9.5 Validation

The tag name is validated according the the rules described above: except for the upper case tags indicated in the GFA1 specification, and the TS header tag, all other tags must contain at least one lower case letter.

```
"VN:i:1"  # => in header: allowed, elsewhere: error
"TS:i:1"  # => allowed in headers and GFA2 Edges
"KC:i:1"  # => allowed in links, containments, GFA1/GFA2 segments
"xx:i:1"  # => custom tag, always allowed
"xxx:i:1" # => error: name is too long
"x:i:1"   # => error: name is too short
"11:i:1"  # => error: at least one letter must be present
```

The datatype must be one of the datatypes specified above. For predefined tags, Gfapy also checks that the datatype given in the specification is used.

```
"xx:X:1" # => error: datatype X is unknown
"VN:i:1" # => error: VN must be of type Z
```

The data must be a correctly formatted string for the specified datatype or a Python object whose string representation is a correctly formatted string.

```
# current value: xx:i:2
line.xx = 1   # OK
line.xx = "1" # OK, value is set to 1
line.xx = "A" # error
```

Depending on the validation level, more or less checks are done automatically (see validation chapter). Per default - validation level (1) - validation is performed only during parsing or accessing values the first time, therefore the user must perform a manual validation if he changes values to something which is not guaranteed to be correct. To trigger a manual validation, the user can call the method `validate_field(fieldname)` to validate a single tag, or `validate()` to validate the whole line, including all tags.

```
line.xx = "A"
line.validate_field("xx") # validates xx
# or, to validate the whole line, including tags:
line.validate()
```

9.6 Reading and writing tags

Tags can be read using a property on the Gfapy line object, which is called as the tag (e.g. line.xx). A special version of the property prefixed by `try_get_` raises an error if the tag was not available (e.g. `line.try_get_LN`), while the tag property (e.g. `line.LN`) would return `None` in this case. Setting the value is done assigning a value to it the tag name method (e.g. `line.TS = 120`). In alternative, the `set(fieldname, value)`, `get(fieldname)` and `try_get(fieldname)` methods can also be used. To remove a tag from a line, use the `delete(fieldname)` method, or set its value to `None`.

```
# line is "H xx:i:12"
line.xx  # => 1
line.xy  # => nil
line.try_get_xx    # => 1
line.try_get_xy    # => error: xy is not defined
line.get("xx")     # => 1
line.try_get("xy") # => error, xy is not defined
line.xx = 2        # => value of xx is changed to 2
line.xx = "a"      # => error: not compatible with existing type (i)
line.xy = 2        # => xy is created and set to 2, type is auto-set to i
line.set("xy", 2)  # => sets xy to 2
line.delete("xy")  # => tag is eliminated
line.xx = None     # => tag is eliminated
```

The `tagnames` property of gfapy Line instances is a list of the names (as strings) of all defined tags for a line.

```
print("Line contains the following tags:")
for t in line.tagnames:
  print(t)
if "VN" in line.tagnames:
  # do something with line.VN value
```

When a tag is read, the value is converted into an appropriate object (see Python classes in the datatype table above). When setting a value, the user can specify the value of a tag either as a Python object, or as the string representation of the value.

```
# line is: H xx:i:1 xy:Z:TEXT xz:J:["a","b"]
line.xx # => 1 (Integer)
```

```
line.xy # => "TEXT" (String)
line.xz # => ["a", "b"] (Array)
```

The string representation of a tag can be read using the `field_to_s(fieldname)` method. The default is to only output the content of the field. By setting ‘‘tag: true‘‘‘, the entire tag is output (name, datatype, content, separated by colons). An exception is raised if the field does not exist.

```
# line is: H xx:i:1
line.xx # => 1
line.field_to_s("xx") # => "1"
line.field_to_s("xx", tag=True) # => "xx:i:1"
```

9.7 Datatype of custom tags

The datatype of an existing custom field (but not of predefined fields) can be changed using the `set_datatype(fieldname, datatype)` method. The current datatype specification can be read using `get_datatype(fieldname)`.

```
# line is: H xx:i:1
line.get_datatype("xx") # => "i"
line.set_datatype("xx", "Z")
```

If a new custom tag is specified, Gfapy selects the correct datatype for it: i/f for numeric values, J/B for arrays, J for hashes and Z for strings and strings. If the user wants to specify a different datatype, he may do so by setting it with `set_datatype()` (this can be done also before assigning a value, which is necessary if full validation is active).

```
# line has not tags
line.xx = "1" # => "xx:Z:1" created
line.xx       # => "1"
line.set_datatype("xy", "i")
line.xy = "1" # => "xy:i:1" created
line.xy       # => 1
```

9.8 Arrays of numerical values

`B` and `H` tags represent array with particular constraints (e.g. they can only contain numeric values, and in some cases the values must be in predefined ranges). In order to represent them correctly and allow for validation, Python classes have been defined for both kind of tags: `gfapy.ByteArray` for `H` and `gfapy.NumericArray` for `B` fields.

Both are subclasses of list. Object of the two classes can be created by passing an existing list or the string representation to the class constructor.

```
# create a byte array instance
gfapy.ByteArray([12,3,14])
gfapy.ByteArray("A012FF")
# create a numeric array instance
gfapy.NumericArray("c,12,3,14")
gfapy.NumericArray([12,3,14])
```

Instances of the classes behave as normal lists, except that they provide a #validate() method, which checks the constraints, and that their string representation is the GFA string representation of the field value.

```
gfapy.ByteArray([12,1,"1x"]).validate() # error: 1x is not a valid value
str(gfapy.ByteArray([12,3,14])) # => "c,12,3,14"
```

For numeric values, the `compute_subtype()` method allows to compute the subtype which will be used for the string representation. Unsigned subtypes are used if all values are positive. The smallest possible subtype range is selected. The subtype may change when the range of the elements changes.

```
gfapy.NumericValue([12,13,14]).compute_subtype() # => "C"
```

9.9 Special cases: custom records, headers, comments and virtual lines.

GFA2 allows custom records, introduced by record type strings other than the predefined ones. Gfapy uses a pragmatical approach for identifying tags in custom records, and tries to interpret the rightmost fields as tags, until the first field from the right raises an error; all remaining fields are treated as positional fields.

```
"X a b c xx:i:12" # => xx is tag, a, b, c are positional fields
"Y a b xx:i:12 c" # => all positional fields, as c is not a valid tag
```

For easier access, the entire header of the GFA is summarized in a single line instance. A class (`gfapy.FieldArray`) has been defined to handle the special case when multiple H lines define the same tag (see "Header" chapter for details).

Comment lines are represented by a subclass of the same class (`gfapy.Line`) as the records. However, they cannot contain tags: the entire line is taken as content of the comment. See the "Comments" chapter for more information about comments.

```
"# this is not a tag: xx:i:1" # => xx is not a tag, xx:i:1 is part of the comment
```

Virtual `gfapy.Line` instances (e.g. segment instances automatically created because of not yet resolved references found in edges) cannot be modified by the user, and tags cannot be specified for them. This includes all instances of the `gfapy::Line::Unknown` class. See the "References" chapter for more information about virtual lines.

CHAPTER

TEN

REFERENCES

Some fields in GFA lines contain identifiers or lists of identifiers (sometimes followed by orientation strings), which reference other lines of the GFA file. In Gfapy it is possible to follow these references and traverse the graph.

10.1 Connecting a line to a Gfa object

In stand-alone line instances, the identifiers which reference other lines are either strings containing the line name, pairs of strings (name and orientation) in a `gfapy.OrientedLine` object, or lists of lines names or `gfapy.OrientedLine` objects.

Using the `add_line(line)` (alias: `append(line)`) method of the `gfapy.Gfa` object, or the equivalent `connect(gfa)` method of the gfapy.Line instance, a line is added to a Gfa instance (this is done automatically when a GFA file is parsed). All strings expressing references are then changed into references to the corresponding line objects. The method `is_connected()` allows to determine if a line is connected to an gfapy instance. The read-only property `gfa` contains the `gfapy.Gfa` instance to which the line is connected.

```
link.is_connected() # => False
link.gfa            # => None
link.from_segment   # => "A"
link.connect(gfa)   # or gfa.add_line(link); or gfa.append(link)
link.is_connected() # => True
link.gfa            # => gfapy.Gfa(...)
link.from_segment   # => gfapy.Segment("S\tA\t*", ...)
```

10.2 References for each record type

The following tables describes the references contained in each record type. The notation `[]` represent lists.

10.2.1 GFA1

Record type	Fields	Type of reference
Link	from, to	Segment
Containment	from, to	Segment
Path	segment_names,	[OrientedLine(Segment)]
	links (1)	[OrientedLine(Link)]

(1): paths contain information in the fields segment_names and overlaps, which allow to find the identify from which they depend; these links can be retrieved using `links` (which is not a field).

10.2.2 GFA2

Record type	Fields	Type of reference
Edge	sid1, sid2	Segment
Gap	sid1, sid2	Segment
Fragment	sid	Segment
Set	items	[Edge/Set/Path/Segment]
Path	items	[OrientedLine(Edge/Set/Segment)]

10.3 Backreferences for each record type

When a line containing a reference to another line is connected to a Gfa object, backreferences to it are created in the targeted line.

For each backreference collection a read-only property exist, which is named as the collection (e.g. `dovetails_L` for segments). Note that the reference list returned by these arrays are read-only and editing the references is done using other methods (see the section "Editing reference fields" below).

```
segment.dovetails_L # => [gfapy.line.edge.Link(...), ...]
```

The following tables describe the backreferences collections for each record type.

10.3.1 GFA1

Record type	Backreferences
Segment	dovetails_L
	dovetails_R
	edges_to_contained
	edges_to_containers
	paths
Link	paths

10.3.2 GFA2

Record type	Backreferences	Type
Segment	dovetails_L	E
	dovetails_R	E
	edges_to_contained	E
	edges_to_containers	E
	internals	E
	gaps_L	G
	gaps_R	G
	fragments	F
	paths	O
	sets	U
Edge	paths	O
	sets	U
O Group	paths	O
	sets	U
U Group	sets	U

10.3.3 Segment backreference convenience methods

For segments, additional methods are available which combine in different way the backreferences information. The `dovetails_of_end(end)` and `gaps_of_end(end)` methods take an argument "L" or "R" and return the dovetails overlaps (or gaps) of the left or, respectively, right end of the segment sequence are returned (equivalent to `dovetails_L`/`dovetails_R` and `gaps_L`/`gaps_R`).

The segment `containments` methods returns both containments where the segment is the container or the contained segment. The segment `edges` property is a list of all edges (dovetails, containments and internals) with a reference to the segment.

Other methods directly compute list of segments from the edges lists mentioned above. The `neighbours_L`, `neighbours_R` properties and the ``neighbours(end)`method computes the set of segment instances which are connected by dovetails to the segment. The `segment`containers`and`contained`` properties similarly compute the set of segment instances which, respectively, contains the segment, or are contained in the segment.

```
s.dovetails_of_end("L") # => [gfapy.line.edge.Link(...), ...]
s.dovetails_L == segment.dovetails_of_end("L") # => True
s.gaps_of_end("R") # => []
s.edges # => [gfapy.line.edge.Link(...), ...]
s.neighbours_L # => [gfapy.line.segment.GFA1(...), ...]
s.containers # => [gfapy.line.segment.GFA1(...), ...]
```

10.4 Multiline group definitions

The GFA2 specification opens the possibility (experimental) to define groups on multiple lines, by using the same ID for each line defining the group. This is supported by gfapy.

This means that if multiple `gfapy.line.group.Ordered` or `gfapy.line.group.Unordered` instances connected to a Gfa object have the same `gid`, they are merged into a single instance (technically the last one getting added to the graph object). The items list are merged.

The tags of multiple line defining a group shall not contradict each other (i.e. either are the tag names on different lines defining the group all different, or, if the same tag is present on different lines, the value and datatype must be the same, in which case the multiple definition will be ignored).

```
gfa.add_line("U\tu1\ts1 s2 s3")
[s.name for s in gfa.sets[-1].items] # => ["s1","s2","s3"]
gfa.add_line("U\tu1\t4 5")
[s.name for s in gfa.sets[-1].items] # => ["s1","s2","s3","s4","s5"]
```

10.5 Induced set and captured path

The item list in GFA2 sets and paths may not contain elements which are implicitly involved. For example a path may contain segments, without specifying the edges connecting them, if there is only one such edge. Alternatively a path may contain edges, without explitely indicating the segments. Similarly a set may contain edges, but not the segments refered to in them, or contain segments which are connected by edges, without the edges themselves. Furthermore groups may refer to other groups (set to sets or paths, paths to paths only), which then indirectly contain references to segments and edges.

Gfapy provides methods for the computation of the sets of segments and edges which are implied by an ordered or unordered group. Thereby all references to subgroups are resolved and implicit elements are added, as described in the specification. The computation can, therefore, only be applied to connected lines. For unordered groups, this computation is provided by the method `induced_set()`, which returns an array of segment and edge instances. For ordered group, the computation is provided by the method `captured_path()`, whcih returns a list of `gfapy.OrientedLine` instances, alternating segment and edge instances (and starting and ending in segments).

The methods `induced_segments_set()`, `induced_edges_set()`, `captured_segments()` and `captured_edges()` return, respectively, the list of only segments or edges, in ordered or unordered groups.

```
gfa.add_line("U\tu1\ts1 s2 s3")
u = gfa.sets[-1]
u.induced_edges_set # => [gfapy.line.edge.GFA2("E\te1\ts1+\ts2-...", ...)]
[l.name for l in u.induced_set ] # => ["s1", "s2", "s3", "e1"]
```

10.6 Disconnecting a line from a Gfa object

Lines can be disconnected using the `rm(line)` method of the `gfapy.Gfa` object or the `disconnect()` method of the line instance.

```
line = gfa.segment("sA")
gfa.rm(line)
# or equivalent:
line.disconnect()
```

Disconnecting a line affects other lines as well. Lines which are dependent on the disconnected line are disconnected as well. Any other reference to disconnected lines is removed as well. In the disconnected line, references to lines are transformed back to strings and backreferences are deleted.

The following tables show which dependent lines are disconnected if they refer to a line which is being disconnected.

10.6.1 GFA1

Record type	Dependent lines
Segment	links (+ paths), containments
Link	paths

10.6.2 GFA2

Record type	Dependent lines
Segment	edges, gaps, fragments, sets, paths
Edge	sets, paths
Sets	sets, paths

10.7 Editing reference fields

In connected line instances, it is not allowed to directly change the content of fields containing references to other lines, as this would make the state of the Gfa object invalid.

Besides the fields containing references, some other fields are read-only in connected lines. Changing some of the fields would require moving the backreferences to other collections (position fields of edges and gaps, `from_orient` and `to_orient` of links). The overlaps field of connected links is readonly as it may be necessary to identify the link in paths.

10.7.1 Renaming an element

The name field of a line (e.g. segment `name`/`sid`) is not a reference and thus can be edited also in connected lines. When the name of the line is changed, no manual editing of references (e.g. from/to fields in links) is necessary, as all lines which refer to the line will still refer to the same instance. The references to the instance in the Gfa lines

collections will be automatically updated. Also, the new name will be correctly used when converting to string, such as when the Gfa instance is written to a GFA file.

Renaming a line to a name which already exists has the same effect of adding a line with that name. That is, in most cases, `gfapy.NotUniqueError` is raised. An exception are GFA2 sets and paths: in this case the line will be appended to the existing line with the same name (as described in "Multiline group definitions").

10.7.2 Adding and removing group elements

Elements of GFA2 groups can be added and removed from both connected and non-connected lines, using the following methods.

To add an item to or remove an item from an unordered group, use the methods `add_item(item)` and `rm_item(item)`, which take as argument either a string (identifier) or a line instance.

To append or prepend an item to an ordered group, use the methods `append_item(item)` and `prepend_item(item)`. To remove the first or the last item of an ordered group use the methods `rm_first_item()` and `rm_last_item()`.

10.7.3 Editing read-only fields of connected lines

Editing the read-only information of edges, gaps, links, containments, fragments and paths is more complicated. These lines shall be disconnected before the edit and connected again to the Gfa object after it. Before disconnecting a line, you should check if there are other lines dependent on it (see tables above). If so, you will have to disconnect these lines first, eventually update their fields and reconnect them at the end of the operation.

10.8 Virtual lines

The order of the lines in GFA is not prescribed. Therefore, during parsing, or constructing a Gfa in memory, it is possible that a line is referenced to, before it is added to the Gfa instance. Whenever this happens, Gfapy creates a "virtual" line instance.

Users do not have to handle with virtual lines, if they work with complete and valid GFA files.

Virtual lines are similar to normal line instances, with some limitations (they contain only limited information and it is not allowed to add tags to them). To check if a line is a virtual line, one can use the `is_virtual()` method of the line.

As soon as the parser founds the real line corresponding to a previously introduced virtual line, the virtual line is exchanged with the real line and all references are corrected to point to the real line.

```
g = gfapy.Gfa()
g.add_line("S\t1\t*")
g.add_line("L\t\1\t+\t2\t+\t*")
l = g.dovetails[-1]
g.segment("1").is_virtual() # => False
g.segment("2").is_virtual() # => True
l.to_segment == g.segment("2") # => True
g.segment("2").dovetails = [l] # => True
g.add_line("S\t2\t*")
g.segment("2").is_virtual() # => False
l.to_segment == g.segment("2") # => True
g.segment("2").dovetails = [l] # => True
```

CHAPTER

ELEVEN

THE HEADER

GFA files may contain one or multiple header lines (record type: "H"). These lines may be present in any part of the file, not necessarily at the beginning.

Although the header may consist of multiple lines, its content refers to the whole file. Therefore in Gfapy the header is accessed using a single line instance (accessible by the `header` method). Header lines contain only tags. If not header line is present in the Gfa, then the header line object will be empty (i.e. contain no tags).

Note that header lines cannot be connected to the Gfa as other lines (i.e. calling `connect` on them raises an exception). Instead they must be merged to the existing Gfa header, using `add_line(line)` on the gfa instance.

```
gfapy.Line.from_string("H\tnn:f:1.0").connect(gfa) # exception
gfa.add_line("H\tnn:f:1.0") # this works!
gfa.header.nn # => 1.0
```

11.1 Multiple definitions of the predefined header tags

For the predefined tags (`VN` and `TS`), the presence of multiple values in different lines is an error, unless the value is the same in each instance (in which case the repeated definitions are ignored).

```
gfa.add_line("H\tVN:Z:1.0")
gfa.add_line("H\tVN:Z:1.0") # ignored
gfa.add_line("H\tVN:Z:2.0") # exception!
```

11.2 Multiple definitions of custom header tags

If the tags are present only once in the header in its entirety, the access to the tags is the same as for any other line (see Tags chapter).

However, the specification does not forbid custom tags to be defined with different values in different header lines (which we name "multi-definition tags"). This particular case is handled in the next sections.

11.3 Reading multi-definitions tags

Reading, validating and setting the datatype of multi-definition tags is done using the same methods as for all other lines (see Tags chapter). However, if a tag is defined multiple times on multiple H lines, reading the tag will return a list of the values on the lines. This array is an instance of the subclass `gfapy.FieldArray` of list.

```
gfa.add_line("H\txx:i:1")
gfa.add_line("H\txx:i:2")
gfa.add_line("H\txx:i:3")
gfa.header.xx # => gfapy.FieldArray("i", [1,2,3])
```

11.4 Setting tags

There are two possibilities to set a tag for the header. The first is the normal tag interface (using `set` or the tag name property). The second is to use `add`. The latter supports multi-definition tags, i.e. it adds the value to the previous ones (if any), instead of overwriting them.

```
gfa.header.xx # => None
gfa.header.add("xx", 1)
gfa.header.xx # => 1
gfa.header.add("xx", 2)
gfa.header.xx # => gfapy.FieldArray("i", [1,2])
gfa.header.set("xx", 3)
gfa.header.xx # => 3
```

11.5 Modifying field array values

Field arrays can be modified directly (e.g. adding new values or removing some values). After modification, the user may check if the array values remain compatible with the datatype of the tag using the `validate_field` method.

```
gfa.header.xx # => gfapy.FieldArray([1,2,3])
gfa.header.validate_field("xx") # => True
gfa.header.xx.append("X")
gfa.header.validate_field("xx") # => False
```

If the field array is modified using array methods which return a list or data of any other type, a field array must be constructed, setting its datatype to the value returned by calling `get_datatype(tagname)` on the header.

```
gfa.header.xx # => gfapy.FieldArray([1,2,3])
gfa.header.xx = gfa.FieldArray(gfa.header.get_datatype("xx"),
                               map(lambda x: x+1, gfa.header.xx))
gfa.header.xx # => gfapy.FieldArray([2,3,4])
```

11.6 String representation of the header

For consinstency with other line types, the string representation of the header is a single-line string, eventually non standard-compliant, if it contains multiple instances of the tag. (and when calling `field_to_s(tag)` for a tag present multiple times, the output string will contain the instances of the tag, separated by tabs).

However, when the Gfa is output to file or string, the header is splitted into multiple H lines with single tags, so that standard-compliant GFA is output. The splitted header can be retrieved using the `headers` method on the Gfa instance.

```
gfa.header.field_to_s("xx") # => "xx:i:1\txx:i:2"
str(gfa.header) # => "H\tVN:Z:1.0\txx:i:1\txx:i:2"
[str(h) for h in gfa.headers] # => ["H\tVN:Z:1.0", "H\txx:i:1", "H\txx:i:2"]
str(gfa) # => """
              H VN:Z:1.0
              H xx:i:1
              H xx:i:2
              """
```

CHAPTER

TWELVE

CUSTOM RECORDS

According to the GFA2 specification, each line which starts with a non-standard record type shall be considered an user- or program-specific record.

Gfapy allows to retrieve custom records and access their data using a similar interface to that for the predefined record types. It assumes that custom records consist of tab-separated fields and that the first field is the record type.

Validation of custom records is very limited; therefore, if you work with custom records, you may define your own validation method and call it when you read or write custom record contents.

12.1 Retrieving, adding and deleting custom records

The custom records of a Gfa instance can be retrieved using its `custom_records` property. This returns a list of all custom records, regardless of the record type.

To retrieve only the custom records of a given type use the method `custom_records_of_type(record_type)`.

```
gfa.custom_records
gfa.custom_records_of_type("X")
```

Adding custom records to and removing them from a Gfa instance is similar to any other line. So to delete a custom record, `disconnect()` is called on the instance. To add a custom record line, the instance or its string representation is added using `add_line` on the Gfa instance.

```
gfa.add_line("X\ta\tb")
gfa.custom_records("X")[-1].disconnect()
```

12.2 Tags

As Gfapy cannot know how many positional fields are present when parsing custom records, an heuristic approach is followed, to identify tags. A field resembles a tag if it starts with `tn:d:` where `tn` is a valid tag name and `d` a valid tag datatype (see Tags chapter). The fields are parsed from the last to the first. As soon as a field is found which does not resemble a tag, all remaining fields are considered positionals (even if another field parsed later resembles a tag).

```
gfa.add_line("X\ta\tb\tcc:i:10\tdd:i:100")
x1 = gfa.custom_records("X")[-1]
x1.cc # => 10
x1.dd # => 100
gfa.add_line("X\ta\tb\tcc:i:10\tdd:i:100\te")
x2 = gfa.custom_records("X")[-1]
x1.cc # => None
x1.dd # => None
```

This parsing heuristics has some consequences on validations. Tags with an invalid tag name (such as starting with a number, or with a wrong number of letters), or an invalid tag datatype (wrong letter, or wrong number of letters) are considered positional fields. The only validation available for custom records tags is thus the validation of the content of the tag, which must be valid according to the datatype.

```
gfa.add_line("X\ta\tb\tcc:i:10\tddd:i:100")
x = gfa.custom_records("X")[-1]
x.cc # => None
# (as ddd:i:100) is considered a positional field
```

12.3 Positional fields

The positional fields in a custom record are called `"field1", "field2", ...`. The user can iterate over the positional field names using the array obtained by calling `positional_fieldnames` on the line.

Positional fields are allowed to contain any character (including non-printable characters and spacing characters), except tabs and newlines (as they are structural elements of the line).

Due to the parsing heuristics mentioned in the Tags section above, invalid tags are sometimes wrongly taken as positional fields. Therefore, the user is responsible of validating the number of positional fields.

```
gfa.add_line("X\ta\tb\tcc:i:10\tdd:i:100")
x = gfa.custom_records("X")[-1]
len(x.positional_fieldnames) # => 2
x.positional_fieldnames # => ["a", "b"]
```

12.4 Extensions

The support for custom fields is limited, as Gfapy does not know which and how many fields are there and how shall they be validated. It is possible to create an extension of Gfapy, which defines new record types: this will allow to use these record types in a similar way to the built-in types. However, extending the library requires sligthly more advanced programming than just using the predefined record types.

The manual for writing extensions is provided as Supplementary Information to the manuscript describing Gfapy.

CHAPTER

THIRTEEN

COMMENTS

GFA lines starting with a # symbol are considered comments. In Gfapy comments are represented by instances of `gfapy.line.Comment`. They have a similar interface to other line instances, with some differences, e.g. they do not support tags.

13.1 Accessing the comments

Adding a comment to a `gfapy.Gfa` instance is done similary to other lines, by using the `add_line(line)` method. The comments of a Gfa object can be accessed using the `comments` method. This returns a list of comment line instances. To remove a comment from the Gfa, you need to find the instance in the list, and call `disconnect()` on it.

```
g.add_line("# this is a comment")
[str(c) for c in g.comments] # => ["# this is a comment"]
g.comments[0].disconnect()
g.comments # => []
```

13.2 Accessing the comment content

The content of the comment line, excluding the initial +#+ and eventual initial spacing characters, is included in the field +content+.

The initial spacing characters can be read/changed using the +spacer+ field. The default value is a single space.

```
g.add_line("# this is a comment")
c = g.comments[-1]
g.content # => "this is a comment"
g.spacer # => " "
```

Tags are not supported by comment lines. If the line contains tags, these are nor parsed, but included in the +content+ field. Trying to set tags values raises exceptions.

```
c = gfapy.Line.from_string("# this is not a tag\txx:i:1")
c.content # => "this is not a tag\txx:i:1"
c.xx # => None
c.xx = 1 # raises an exception
```

CHAPTER

FOURTEEN

ERRORS

All exception raised in the library are subclasses of `gfapy.Error`. This means that `except gfapy.Error` catches all library errors.

Different types of errors are defined and are summarized in the following table:

Error	Description	Examples
Version	An unknown or wrong version is specified or implied	"GFA0"; or GFA1 in GFA2 context
Value	The value of an object is invalid	a negative position is used
Type	The wrong type has been used or specified	Z instead of i used for VN tag; Hash for an i tag
Format	The format of an object is wrong	a line does not contain the expected number of fields
NotUnique	Something should be unique but is not	duplicated tag name or line identifier
Inconsistency	Pieces of information collide with each other	length of sequence and LN tag do not match
Runtime	The user tried to do something which is not allowed	editing from/to field in connected links
Argument	Problem with the arguments of a method	wrong number of arguments in dynamically created method
Assertion	Something unexpected happened	there is a bug in the library

Some error types are generic (such as RuntimeError and ArgumentError), and their definition may overlap that of more specific errors (such as ArgumentError, which overlaps ValueError and TypeError). The user should not rely on the type of error alone, but rather take it as an indication. The error message tries to be informative and for this reason often prints information on the internal state of the relevant variables.

Assertion errors are reserved for those situation where something is implied by the programmer (e.g. a value is implied to be positive at a certain point of the code). It the checks fails, an assertion error is raised. The user may report the problem, as this may indicate a bug (unless the user did something he was not supposed to do, such as calling an API private method).

CHAPTER

FIFTEEN

GRAPH OPERATIONS

Graph operations such as linear paths merging, multiplication of segments and other are provided. These operations are similar to those provided by the RGFA library. A description of these operation can be found in the RGFA paper (Gonnella and Kurtz, 2016).

CHAPTER

SIXTEEN

INDICES AND TABLES

- genindex
- modindex
- search