Biochemfusion

Protein Line Notation

Specification

version 1.0

2008-11-28

Table of Contents

Introduction

The Biochemfusion Protein Line Notation format is a compact text representation of a protein that includes chemically significant annotations.

Although the name implies a single line of text the text may be broken into an arbitrary number of lines to ease e-mail transmission and enhance readability.

Line notation regions

The line notation consists of two distinct regions: the Sequence region and the Properties region:

```
H-ASDF-OH.H-CGTY-OH name="Simple protein" id=P00001
<--- Sequence ----><--- Properties --------------->
```

The sequence region

The Sequence region may contain any number of linefeeds but *cannot* contain white space. Once a white space is encountered the Properties region is assumed to start.

The Sequence region contains chains delimited by periods, '.'. A chain consists of an N-terminal specification, followed by a hyphen, '-', followed by a list of residues, followed by a hyphen, followed by a C-terminal specification.

Terminals

Standard unmodified terminals must be written as "H" (N-terminal) or "OH" (C-terminal).

Modified terminals must be written as the modification name in square brackets, e.g. "[biotin]". Terminal modification names must follow the format and constraints detailed in the section "Modification names". The mechanism for mapping modification names to terminal structures will be determined by the actual implementation.

A final variant is terminals that are the endpoints of a cyclization, in which case the terminal must be written as "(<cycle identifier>)". The format of cyclizations is described in the section "Cyclizations".

Residues

The unmodified residues in a chain are single letter uppercase codes for the unmodified natural amino acids. A small peptide that in IUPAC notation[1] would be

1 IUPAC "Nomenclature and Symbolism for Amino Acids and Peptides" http://www.chem.qmul.ac.uk/iupac/AminoAcid/A1819.html#AA191


```
Ala-Cys-Asp-Glu-Phe-Gly
```

would be written as

```
H-ACDEFG-OH
```

in Protein Line Notation.

Non-standard residues

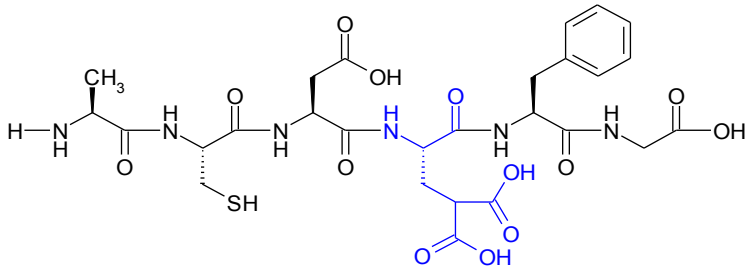
Post-translationally or chemically modified residues are written as the modified residue name in square brackets, e.g. "[4-carboxyglutamate]" and the modified name replaces the single letter code. A Glu-carboxy variant of

```
H-ACDEFG-OH
```

would become

```
H-ACD[4-carboxyglutamate]FG-OH
```

The modified sequence above corresponds to the following chemical structure where the modified residue has been highlighted in blue.



How modification names are mapped to actual structures will be implementation-dependent.

D-form residues

All residues can be transformed into their D-forms by prefixing the residue name (one-letter code or modification name) with “{d}”, e.g.

```
H-AS{d}EF-OH
```

has a D-Glu in position 3.

Bridges and cycles

Bridges and cycles are indicated by identifiers in parenthesis that are added as residue suffixes.

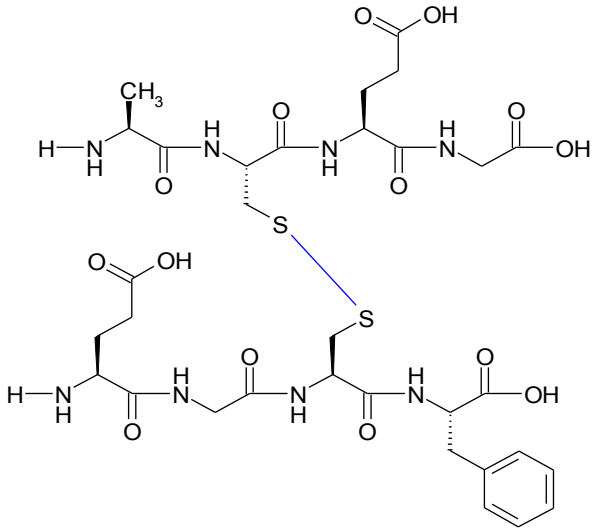
Disulfide bridges

Disulfide bridges are indicated by pure-numeric identifiers that may follow only

cysteine residues. An example is

```
H-AC(1)EG-OH.H-EGC(1)F-OH
```

which corresponds to the following chemical structure where the disulfide bond has been highlighted in blue.



The identifiers chosen do not have to be consecutive or follow any numeric sequence; they only serve to uniquely identify the bridge. This means that the following notation is equivalent to the one above.

```
H-AC(5)EG-OH.H-EGC(5)F-OH
```

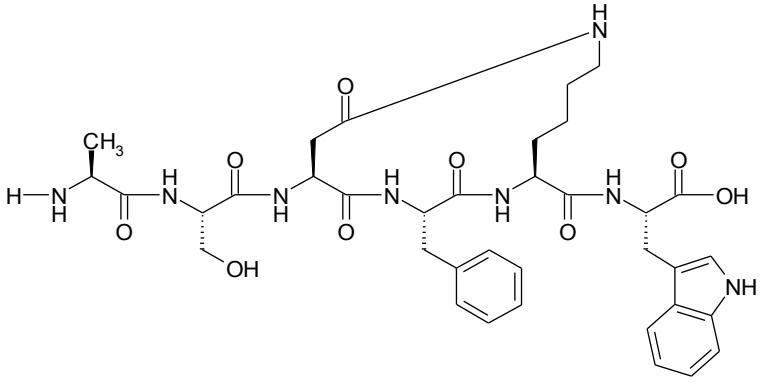
Cyclizations

A cyclization is specified by using the keyword “cyclo” followed by a number that uniquely identifies the particular cycle. A cyclization may only form between a reactive amino group and a reactive acid group. How the actual reaction sites are found will be implementation-dependent.

Cyclizations may form between sidechains, e.g.

```
H-ASD(cyclo1)FK(cyclo1)W-OH
```

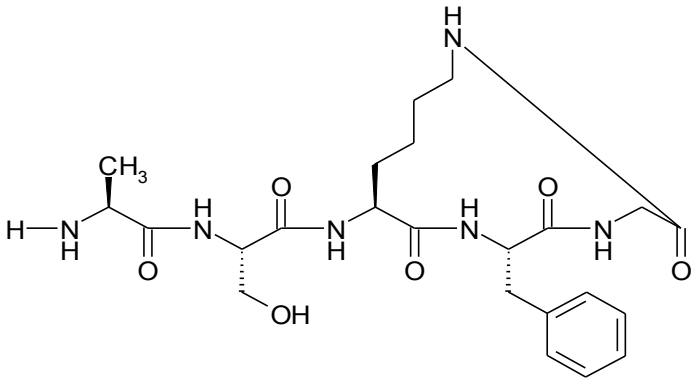
which corresponds to the chemical structure:



Cyclizations may also form between terminals and side chains, e.g.

```
H-ASK(cyclo1)FG-(cyclo1)
```

corresponding to



or between terminals as in the fully cyclic peptide below

```
(cyclo1)-ASDEF-(cyclo1)
```

The numbering follows the same rules as the numbering of disulfide bridges. The numbering does not have to be consecutive, it only serves to uniquely identify a particular cycle.

Modification names

Allowed characters in modification names are

- All alpha-numeric characters `a..z, A..Z, 0..9`
- Square brackets and normal parentheses `[]()`
- Comma, period, hyphen `,.-`
- Apostrophe and underscore `'_`

If square brackets and parentheses are used the brackets and parentheses must be correctly paired within the modification name.

Modification names cannot end with a period.

The Properties region

The Properties region lists properties that are key-value pairs separated by equal signs, '='. Keys, values, and '=' characters may be delimited by any number of white spaces and linefeeds. Linefeeds are ignored as they are in the Sequence region.

Property keys

A property key name is all lowercase. Valid property keys are

Key	Key description
name	Protein name
id	Protein id

Property values

A property value may only contain white space or double quotes if it is surrounded by double quotes. A double quote must be encoded as two consecutive double quotes, e.g.

Value	Correctly quoted
Simple protein	"Simple protein"
Not so "small" protein	"Not so ""small"" protein"
"Nice"-protein	"""Nice""-protein"

Note: The value of the "id" property may contain alphanumeric and underscore characters only - no white space allowed.

