

Biochemfusion

Protein Line Notation

Specification

version 1.1

2009-10-14

Copyright © 2008-2009 Biochemfusion. All rights reserved.

Permission is granted to freely redistribute and copy this document in any form subject to the following restrictions:

- *The document may not be changed in any way.*
- *The document must be distributed in its entirety.*
- *The above copyright notice and these copy restrictions must be kept clearly legible.*

biochemfusion

Table of Contents

Line notation regions	3
The sequence region	3
Terminals	3
Residues	4
Non-standard residues	4
D-form residues	5
Bridges and cycles	5
Disulfide bridges	5
Cyclizations	6
Modification names	7
The Properties region	7
Property keys	7
Property values	7
The end-of-entry region	8
References	8

Changes since version 1.0

Optional end-of-entry region.	The use of the optional end-of-entry delimiter is now described in the sections "Line notation regions" and "The end-of-entry region".
Support for di-selenium bridges.	Disulfide bridges may also form between selenocysteines - see the section "Disulfide bridges".
"References" section added.	External references have been moved from footnotes to a proper section, "References".

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 three distinct regions: the Sequence region, the Properties region, and the optional end-of-entry region:

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

The sequence region

The Sequence region cannot contain white space. Once a white space is encountered the Properties region is assumed to start. The region may contain linefeeds but all linefeeds must be ignored.

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 terminal structure name (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 terminal may also be the endpoint of a cyclization, in which case the terminal must be written as "(<cycle identifier>)". The syntax and representation of cyclizations is described in the section "Cyclizations".

Residues

The unmodified residues in a chain are single letter uppercase residue codes. ¹
A small peptide that in IUPAC notation would be ²

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 indicated by replacing the modified residue's single letter code by a residue structure name (modification name) in square brackets.

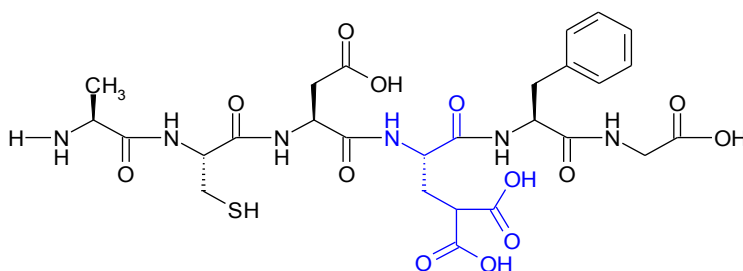
An example would be that a carboxyl variant of a Glu residue is described by the modification name "4-carboxyglutamate". A chain containing a Glu/E residue

H-ACDEFG-OH

would then have a carboxyl variant by

H-ACD[4-carboxyglutamate]FG-OH

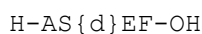
The modified sequence above corresponds to the following chemical structure where the modified residue has been highlighted in blue.



How modification names are mapped to actual residue 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.



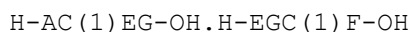
has a D-Glu in position 3.

Bridges and cycles

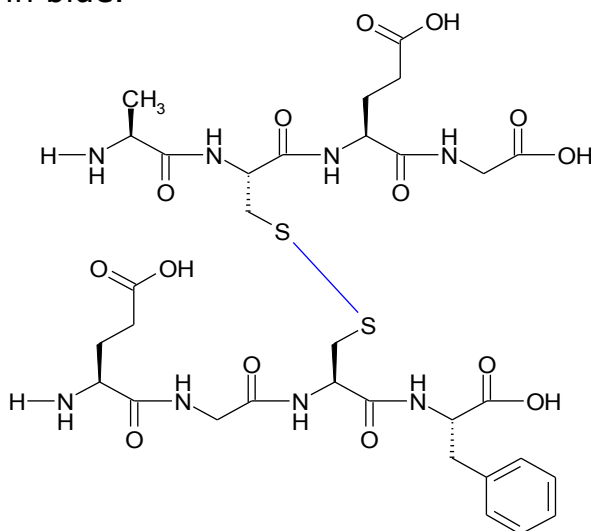
Bridges and cycles may form between residue pairs, terminal pairs, and residue-terminal pairs. Each bridge or cycle endpoint is indicated by a parenthesis pair containing a unique identifier.

Disulfide bridges

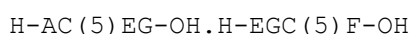
Disulfide bridges are indicated by pure-numeric identifiers that may suffix cysteine or selenocysteine residues. An example is



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 example is equivalent to the one above.



Cyclizations

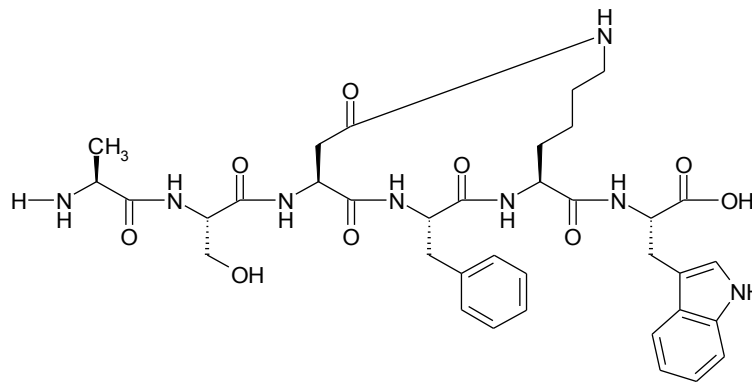
A cyclization differs syntactically from a disulfide bridge by the form of its identifier. A cyclization identifier is formed by the keyword "cyclo" followed by a unique integer.

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 residue sidechains, e.g.

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

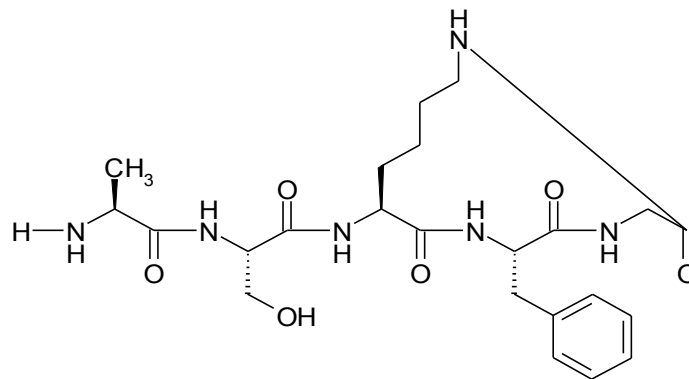
which corresponds to the chemical structure:



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 obeys 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 within a modification name the brackets and parentheses must be correctly paired.

Although periods are allowed within a 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 should be ignored as 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 name value may only contain white space or double quote characters if it is enclosed in double quotes. A double quote within a name value must be encoded as two consecutive double quotes, e.g.

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

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

The end-of-entry region

An optional end-of-entry delimiter may be used to separate a PLN entry from plain text that follows or other PLN entries in a multi-entry text. If no other text follows the PLN entry an end-of-entry delimiter is not required.

An end-of-entry delimiter consists of at least one white space followed by two asterisk '*' characters.

Examples of correct usage:

PLN text	Comments
H-ACDEFG-OH ** which is a fascinating entry...	End-of-entry used to separate PLN from plain text.
H-ACDEFG-OH name=1st_entry**H-QWER-OH name=2nd_entry**H-EFTYS-OH name=final_entry	Three PLN entries. Note that the last entry does not need an end-of-entry delimiter.
H-ACDEFG-OH name="Two stars **" ** which is...	The protein name (here "Two stars **") can contain an end-of-entry delimiter as long as it is correctly quoted.

Examples of invalid end-of-entry delimiter usage:

PLN text	Comments
H-ACDEFG-OH**	Missing white space before '**'.
H-ACDEFG-OH↓ **	Missing white space before '**' (linefeeds are ignored).
H-ACDEFG-OH name="1st_entry **"...and I am quoting here	Incorrectly quoted protein name masks end-of-entry delimiter.

References

- (1) IUPAC "Nomenclature and Symbolism for Amino Acids and Peptides"
<http://www.chem.qmul.ac.uk/iupac/AminoAcid/AA1n2.html#AA1>
- (2) IUPAC "Nomenclature and Symbolism for Amino Acids and Peptides"
<http://www.chem.qmul.ac.uk/iupac/AminoAcid/A1819.html#AA191>