PROTEIN CONSTRUCT NAMING CONVENTION

Introduction

Whilst there are common nomenclature conventions for genes [1] [2] [3] [4] and for proteins themselves [5] the naming of recombinant protein constructs in the scientific literature is exceptionally diverse with a variety of approaches displayed by researchers.

This protocol is given as a set of guidelines (rather than rules) that are intended to make protein construct naming as consistent as possible within a single organisation at least.

Summary of different protein names

These are the definitions of the various protein names that are commonly used:-

Common Protein Name: The commonly used name or abbreviation. e.g. MMP9, cMet or Fibronectin.

Full Protein Name: The full name. e.g. Hepatocyte Growth Factor or Matrix Metalloproteinase 9.

Construct Name: The name of the protein construct as cloned in the vector prior to purification and any post translational or in process modifications. This name should include any pre sequence, mutations, tags, cleavage sites etc. e.g. Pre-Pro-YAP1(51-345)-TEV-6His.

See below for details on the naming convention.

Final Protein Product: This is the name of the construct as finally purified and supplied after any post translational and/or in process modifications, tag cleavage etc. e.g. MMP9(1-100)-Biotin.

Protein Construct Naming Convention

When discussing a protein, where possible the Uniprot [6] reference should be quoted. This shows which species the protein in question derives from. However, ambiguity is still possible. The common name or abbreviation could well refer to more than one form of the protein; isoforms or sequence variants, engineered constructs, activated versus precursor forms or mutant versions perhaps.

To define the exact nature of the protein, construct the following naming convention is proposed. A summary table is shown below, followed by a detailed description of the naming guidelines for the various features a protein construct may contain.

Compile the various features of the protein construct that are relevant to your work using the guidelines given in the table below. Then assemble them in the order that they occur in the protein sequence. For complete unambiguity the actual amino acid sequence of the construct should be included in the paper or report.

A worked example is given at the end of this protocol.

Feature	Convention	Example(s)
Protein	Use common abbreviation or name	YAK1 Yet Another Kinase 1
Species	Define the species	Bovine
Amino acid numbering	Use Uniprot numbering	(33 – 492)
	Can include amino acids for additional clarity	(A33 – G492) or (Ala33 – Gly492)
Pre / Pro / Signal and other similar sequences	If features such as signal, pre- and pro- sequences are to be highlighted place a tag descriptor for the type of sequence followed by the relevant amino acid numbering in round brackets.	pre(1-20)-pro(21-33)-mature YAP1(34-456).
Tags	Use common abbreviations for tags with a hyphen either before or after.	N term: FLAG-, 6His-, GST-, C term: -6His, -AVI
	Multiple tags are added together	6His-cmyc-
Protease cleavage sites	Enter either with common protease abbreviation with a hyphen or with the protease cleavage amino acid sequence	-TEV- Or -ENLYFQG-
Single mutants	In square brackets, show the position number of amino acid that has been changed. In front put the original residue and after put what it has been mutated to	[S345A]

Feature	Convention	Example(s)
Double or Triple mutants	Separated by a comma	[S345A, S383A]
Additions	To N or C terminus add on amino acid sequence followed / preceded by hyphen	Gly-Ser- Or GS-
	Of a sequence from another protein. Indicate species, source and length. Or include amino acid sequence	Apis mel. Melittin sig. (1-21) Represents secretion signal sequence from Apis mellifera (Honeybee) Melittin. Or MKFLVNVALVFMVVYISYIYA-
	For junk amino acids from cloning artefact or other, the letter j may be added preceding a number corresponding to the length of the addition	j23-MMP9 indicates 23 amino acids of junk sequence N terminal to the MMP9 sequence
Insertions	Sequence numbering is used to show the insertion site. Single letter code is used to show inserted sequence, preceded by a double colon ::	MMP9(1-162 ::HHGYFG, 163-495)
Internal Deletions	Use numbering with comma Or Use delta sign D	YAK1(1-122, 136-345) Or YAK1(1-345, D123 – 135)
Multiple Chains	A letter or other symbol defining the chains is added in front of the chain residue numbers. The numbers of the chains are enclosed in round brackets and are separated by a forward slash	Human Insulin (a90-110/b25-54)
Non covalent complexes	Components separated by colon :	CDK4(1-556):CyclinD1(1-224) Compound A23456:MMP9(1-409)
Modifications	Either in process or post translational modifications. Added at the end of the construct name with a hyphen	CDK4(1-556)-biotin CDK4(1-556)-phosphorylated
Labelling	Added in square brackets at the front of the construct name	[N^{15}]MMP9(107-216,391-443) [C^{13},N^{15}]MMP9(107-216,391-443)

Abbreviation or full common protein name:

Use the common abbreviation or name as the core to the protein construct. Where possible the Uniprot [6] reference should be quoted in the report or document. This will define which species the protein derives from. Use the name that appears to be preferred in the literature but mention any alternative names that it may have.

Example: Cathepsin K or Cat K
Uniprot entry P43235 is human Cat K
Also known as Cathepsin O, Cathepsin O2, Cathepsin X

Species:

Define the species that the protein in question derives from.

Example: Janus Kinase 1 or JAK1
Uniprot entry P52332 is murine JAK1

Numbering:

Numbering should be according to the Uniprot [6] entry. Each section of the protein is numbered from the N terminal to the C terminal amino acids contained in round brackets. Hyphens are used to seperate the numbers between these two positions. There is no gap between the protein abbreviation and the first bracket.

For additional clarity you can include the amino acids alongside their positional numbers.

Examples murine JAK1(1-1153) or murine JAK1(M1-K1153) or murine JAK1(Met1-Lys1153)

Signal / Pre- / Pro- and other similar sequences:

If features such as signal, pre- and pro- sequences are to be highlighted place a tag descriptor for the type of sequence followed by the relevant amino acid numbering in round brackets.

Example pre(1-20)-pro(21-33)-mature YAP1(34-456)

Tags:

Protein tags are entered as their common abbreviation with a hyphen e.g. GST-MMP9. Multiple tags have hypens between them e.g. 6His-cmyc-IKK1

Common N term tags: GST- MBP- 6His- FLAG- cmyc- Avi-

Common C term tags: -6His -FLAG -cmyc -Avi

Protease cleavage sites:

Protease cleavage sites are entered either as their common abbreviation with a hyphen e.g. 6His-PS-MMP9 for a PreScission site or as their amino acid sequence with a hyphen e.g. 6His-LVPRGS-MMP9 for a thrombin site.

Common proteases: PS = PreScission, Th = Thrombin, Xa = Factor Xa, TEV = TEV protease, EK = Entrokinase.

Mutations:

Single letter nomenclature in upper case is used to indicate the amino acid change with the position of the changed amino acid included. In square brackets, show the position number of amino acid that has been changed. In front put the original residue and after put what it has been mutated to. Multiple mutations are separated by a comma e.g. [R173Q, G224Y]. They should directly precede the protein abbreviation e.g. [R173Q, G224Y]MMP9(1-404). Note that no gap or hyphen is included between the last square bracket and the protein name.

Additions:

For additions at either termini, the 1 or 3 letter code may be used to indicate the addition e.g. Gly-Ser-MCP1(1-76) or GS-MCP1(9-76) are both correct.

Where an addition is from an alternative source an indication of the origin of the sequence (species and protein) should be shown where possible, also giving the residue numbers added.

Example Apis mel. Melittin sig. (1-21)

Represents secretion signal sequence from Apis mellifera (Honeybee) Melittin.

Alternatively just the added amino acid sequence could be used

MKFLVNVALVFMVVYISYIYA-

For the addition of ‘junk’ sequences as an artifact of cloning the letter j may be added preceding a number corresponding to the length of the addition e.g. j23-MMP9 indicates 23 amino acids of junk sequence N terminal to the MMP9 sequence.

For the addition of amino acids as a result from Gateway cloning the letters gw in lower case may be added.

Insertions:

Sequence numbering is used to show the insertion site then either three letter (for short sequences) or the single letter code is used to detail the inserted sequence, preceded by a double colon ::

Examples MMP9(1-162 ::Gly, 163-495)

MMP9(1-162 ::HHGYFG, 163-495)

Internal Deletions:

For deletion of an internal portion of protein either the remaining fragments should be separated by a comma or a delta sign D can be used.

Examples MMP9(1-107,224-409) or MMP9(1-409 D108-223)

Truncation of Sequences:

As pointed out above numbering is added as (1-x) therefore truncations should be annotated as follows e.g. MCP1(9-76) shows the first 8 amino acids are not included.

Multiple Chain Proteins:

Where multiple chains are present a letter or other symbol defining the chains is added in front of the chain residue numbers. The numbers of the chains are enclosed in round brackets and are separated by a forward slash.

Example Human Insulin (a90-110/b25-54)

Noncovalent Complexes:

These may include copurified proteins or the association with for example a inhibitor compound. Where this is the case the separate molecules should be separated by a colon

Examples Compound A23456:MMP9(1-409)
CDK4(1-556):CyclinD1(1-224).

Modifications:

These may include in process modifications or post translational modifications and are added at the end of the construct name with a hyphen. e.g. the use of NHS-Biotin for modification of Lys side chains is CDK4(1-556)-biotin

If it is helpful to describe the exact details of a site specific modifications the modified amino acid can be included in square brackets annotated to indicate the modification that has occurred.

Example CDK4(1-556)-[phospho Y200,Y206]

Labelled Protein:

Labels should be placed in square brackets at the beginning of the name with each label indicated clearly (i.e. just stating double or triple labelled is not all that useful)

Examples [N^{15}]MMP9(107-216,391-443)
[C^{13},N^{15}]MMP9(107-216,391-443)

Example

MMP9 is protease with a common name of Matrix metalloproteinase-9. The Uniprot entry for the human version is P14780 and it is also known as 92kDa gelatinase, 92kDa type IV collagenase or gelatinase B.

It is 707 amino acids long

1 – 19 is a signal peptide
20 – 93 is a pro-peptide
107 – 707 is main active MMP9 chain

We have cloned a construct with just the pro-peptide and the main chain preceded by a 6His tag that has a TEV cleavage site after it. The construct is to be transfected into HEK293 cells so we have also added a signal sequence from honey bee melittin to promote secretion into the media during culture. The cloned construct name would be:-

Apis mel. Melittin sig. (1-21)-6His-TEV-Pro(20 – 93)-MMP9(107 – 707)

Or if we didn’t need to pass on detail about the signal sequence and the pro domain we could just use this simpler construct name which would also be correct:-

6His-TEV-MMP9(20 – 707)

During purification we cleave the tag off using TEV and perform an autoactivation step that releases the mature active MMP9. Because we expressed it in HEK cells we have managed to achieve glycosylation at positions 120 and 127. Therefore the final purified protein name would be:-

MMP9(107 – 707)-[N glyco Asn120, Asn127]

Again, we could leave off the information about the glycosylation if that was not really needed:-

MMP9(107-707)

References

[1] HUGO Gene Nomenclature Committee, “The resource for approved human gene nomenclature,” [Online]. Available: https://www.genenames.org/.

[2] den Dunnen J T et al, “HGVS Recommendations for the Description of Sequence Variants: 2016 Update,” *Human Mutation,* vol. 37, no. 6, pp. 564 - 569, 2016.

[3] Human Genome Variation Society, “Sequence Variant Nomenclature,” [Online]. Available: http://varnomen.hgvs.org/.

[4] Wikepedia, “Gene Nomenclature,” [Online]. Available: https://en.wikipedia.org/wiki/Gene_nomenclature.

[5] NCBI, “International Protein Nomenclature Guidelines,” [Online]. Available: https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/.

[6] UniProt Consortium, “UniProt,” [Online]. Available: https://www.uniprot.org/.