Current output from neXtProt (https://api.nextprot.org/export/entry/NX\_P01308.peff):

# PEFF 1.0

# //

# DbName=neXtProt: NX\_P01308

# DbSource=https://www.nextprot.org

# DbVersion=2017-08-01

# Prefix=nxp

# NumberOfEntries=1

# SequenceType=AA

# GeneralComment=Copyrighted by the SIB Swiss Institute of Bioinformatics

# GeneralComment=Distributed under the Creative Commons Attribution-NoDerivs License

# //

>nxp:NX\_P01308-1 \DbUniqueId=NX\_P01308-1 \PName=Insulin isoform Iso 1 \GName=INS \NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=110 \SV=1 \EV=228 \PE=1 \ModRes=(31||Disulfide)(96||Disulfide) (43||Disulfide)(109||Disulfide)(95||Disulfide)(100||Disulfide) \VariantSimple=(2|T)

(6|C)(6|G)(6|H)(8|Q)(9|S)(12|V)(18|R)(21|L)(22|V)(23|S)(23|T)(24|D)(24|V)(29|D)(29|P)(32|R)(32|S)(34|D)(35|P)(38|V)(42|A)(43|G)(44|R)(45|K)(46|Q)(47|V)(48|C)(48|S)(49|L)(51|I)(52|R)(53|E)(53|T)(55|C)(55|H)(56|W)(58|V)(63|A)(63|L)(64|W)(65|L)(68|M)(70|R)(71|V)(73|C)(75|D)(76|N)(76|R)(79|L)(81|V)(83|K)(84|R)(85|Y)(89|C)(89|H)(89|L)(89|P)(90|C)(90|D)(92|L)(93|K)(94|K)(96|S)(96|Y)(98|R)(101|C)(103|C)(106|D)(108|C) \Processed=(1|24|signal peptide)(25|54|mature protein)(57|87|maturation peptide)(90|110|mature protein)

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED

LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

Proposed proteoform format to represent Insulin:

# PEFF 1.0

# //

# DbName=neXtProt: NX\_P01308

# DbSource=https://www.nextprot.org

# DbVersion=2017-08-01

# HasAnnotationIdentifiers=true

# Prefix=nxp

# NumberOfEntries=1

# SequenceType=AA

# GeneralComment=Copyrighted by the SIB Swiss Institute of Bioinformatics

# GeneralComment=Distributed under the Creative Commons Attribution-NoDerivs License

# //

>nxp:NX\_P01308-1 \DbUniqueId=NX\_P01308-1 \PName=Insulin isoform Iso 1 \GName=INS \NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=110 \SV=1 \EV=228 \PE=1 \ModResPsi=(0:53|MOD:00087|N6-myristoyl-L-lysine) \ModRes=(1:31||Disulfide)(2:96||Disulfide) (3:43||Disulfide)(4:109||Disulfide)(5:95||Disulfide)(6:100||Disulfide) \VariantSimple=(7:2|T) (8:6|C)(9:6|G)(10:6|H)(11:8|Q)(12:9|S)(13:12|V)(14:18|R)(15:21|L)(16:22|V)(17:23|S)(18:23|T)(19:24|D)(20:24|V)(21:29|D)(22:29|P)(23:32|R)(24:32|S)(25:34|D)(26:35|P)(27:38|V)(28:42|A)(29:43|G)(30:44|R)(31:45|K)(32:46|Q)(33:47|V)(34:48|C)(35:48|S)(36:49|L)(37:51|I)(38:52|R)(39:53|E)(40:53|T)(41:55|C)(42:55|H)(43:56|W)(44:58|V)(45:63|A)(46:63|L)(47:64|W)(48:65|L)(49:68|M)(50:70|R)(51:71|V)(52:73|C)(53:75|D)(54:76|N)(55:76|R)(56:79|L)(57:81|V)(58:83|K)(59:84|R)(60:85|Y)(61:89|C)(62:89|H)(63:89|L)(64:89|P)(65:90|C)(66:90|D)(67:92|L)(68:93|K)(69:94|K)(70:96|S)(71:96|Y)(72:98|R)(73:101|C)(74:103|C)(75:106|D)(76:108|C) \Processed=(77:1|24|signal peptide)(78:25|54|mature protein)(79:57|87|maturation peptide)(80:90|110|mature protein) \DisulfideBond=(81:1,2|between chains)(82:3,4|between chains)(83:5,6|A chain only) \Proteoform=(NX\_P01308-1-pf1|1-110||preproinsulin)(NX\_P01308-1-pf2|25-110||proinsulin)(NX\_P01308-1-pf3|25-110|1,2,3,4,5,6|proinsulin with disulfide mods)(NX\_P01308-1-pf4|90-110||Insulin A chain cleaved)(NX\_P01308-1-pf5|90-110|3,4,5,6|Insulin A chain modified)(NX\_P01308-1-pf6|25-54||Insulin B chain cleaved)(NX\_P01308-1-pf7|25-54|5,6|Insulin B chain cleaved)(NX\_P01308-1-pf8|25-53|0,1,3|B chain in an extracellular region)(NX\_P01308-1-pf9|57-87||C peptide cleaved)(NX\_P01308-1-pf10|57-87||C peptide cleaved)(NX\_P01308-1-pf11|90-110,25-54|81,82,83|Insulin: chains A and B joined)

MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED

LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

Notes:

- I added MOD:00087 manually since it was in Francisco’s example, but not in the neXtProt output. I wonder if it should be in neXtProt?

- The AnnotationIdentifiers flag at the top lets the parser know to expect extended features (the identifiers)

- Still just one entry, using all the usual PEFF conventions. From a naïve parser (e.g. shotgun search engine) point of view, all that’s different is that there are now index: constructs (in green) to deal with (parse correctly but discard) and then some additional keywords (\DisulfideBond and \Proteoform) that it will not know how to deal with but can easily skip.

- Showing how disulfide bonds could be linked with new keyword with format: \DisulfideBond=(newId:id1,id2|optionalTag)

- \Proteoform showing a list of proteoforms of interest from Francisco. New \Proteoform keyword has suggested format: \Proteoform=(proteoformUniqueId|start-stop|variationList|optionalTag). I suggest that within a file, all proteoform ids must be unique; I arbitrarily used the protein accession -pf1, -pf2, -pf3, etc. VariationList is nominally just a comma-separated list of elements in the file. The start-stop is just within the main sequence entry. For the chain A+B, I turned this into a comma-separated list of two ranges. There is potentially additional fanciness that could happen here for those interested in this.

- Although a completely separate piece of work, I can easily see someone writing a simple little indexer as a sidecar file to a PEFF file that indexes all the entry accessions and all the proteoform accessions for easy random access.

Question for Lydie: what is the reason that neXtProt has:

\ModRes=(31||Disulfide)(96||Disulfide)(43||Disulfide)(109||Disulfide)(95||Disulfide)(100||Disulfide)

Instead of:

\ModResPsi=(31|MOD:00789|half cystine)(96|MOD:00789|half cystine)(43| MOD:00789|half cystine)(109|MOD:00789|Disulfide)(95|MOD:00789|half cystine)(100|MOD:00789|half cystine)

Should this change?

<https://www.ebi.ac.uk/ols/ontologies/mod/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMOD_00798>

Responses to Francisco:

Francisco said:

// Insulin: chains B and A join with a disulfide bond. C peptide is free

How to represent this?

P01308,90-110 -- SS -- P01308,25-54 + P01308,57-87

Eric: I would say that all you need is “P01308,90-110 -- SS -- P01308,25-54”. If the C peptide is free then it is a different proteoform, right?

I would say this case is modeled by NX\_P01308-1-pf11 with a separate NX\_P01308-1-pf9 in the above encoding.

Does that seem right?

\* The "-1" was omitted to refer to the default isoform: P01308-1

Eric: I kept it in just because it seemed better for this example. Up to the writer.

\* The coordinates vary +- 5 positions. Some authors say Proinsulin has 81

residues, some say 85.

\* Cleavage sites for C peptide also differ sometimes

Eric: Authors of papers may differ, but I think this PEFF extension lets authors of PEFF files be exact on what they mean!