Determination of the lexicographical order of protein chains based on SPL Index information.

Amino acid (AA ) sequence is represented by single letter notation. **I would not recommend splitting AA sequence into substrings and using InChI for substrings because such algorithm would be poorly reproducible and the hash wouldn’t be computable from SPL Index information alone.**

Lexicographical order is determined in two iterations

1. First iteration

Each modified AA is replaced by combination of its InChI and its connection points. All the info is explicitly present in the SPL Index file. For example, in file bf170df7-8153-45c1-9a21-6e1aa1b4c97e.xml,

amino acid SU3\_1 can be replaced by the InChI of the pyroglutamic acid: InChI=1S/C5H7NO3/c7-4-2-1-3(6-4)5(8)9/h3H,1-2H2,(H,6,7)(H,8,9)/t3-/m0/s1 combined with connection atoms [6,5]. The resultant string may be represented as [ InChI=1S/C5H7NO3/c7-4-2-1-3(6-4)5(8)9/h3H,1-2H2,(H,6,7)(H,8,9)/t3-/m0/s1]|[6,5] or likewise.

Each cysteine that forms a disulfide bond can be replaced by the InChI of the disulfide linker: InChI=1S/C6H12N2O4S2/c7-3(5(9)10)1-13-14-2-4(8)6(11)12/h3-4H,1-2,7-8H2,(H,9,10)(H,11,12)/t3-,4-/m0/s1 combined with connection atoms: (7,5;8,6). The resultant string may be represented as [InChI=1S/C6H12N2O4S2/c7-3(5(9)10)1-13-14-2-4(8)6(11)12/h3-4H,1-2,7-8H2,(H,9,10)(H,11,12)/t3-,4-/m0/s1 ]|[7,5;8,6] or likewise.

After concatenation of single letter notation and the strings that represent modified AAs, the AA sequences can be ordered in a lexicographical order and assigned ranks. It is possible that two or more sequences have the same rank.

1. Second iteration

Each modification of type “link” is extended by adding the rank of the aa sequence and position of other amino acids that form that link. If the second aa belongs to the same sequence the rank is skipped.

After concatenation of single letter notation and the extended strings that represent modified AAs, the AA sequences can be ordered in a lexicographical order and assigned NEW ranks.

1. Next Iteration

Replace sequence ranks from Iteration 1 with sequence ranks from Iteration 2. Order lexicographically and re-rank. Iterations stop when no new ranks can be assigned.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | X | A | A | A | A | X |  |  |  |  | A | A | X | A | A | A | A | X |  |  |

Iteration1: RE-RANKING

Seq|its rank|linker:rank of the distal seq : position on this-position on distal;…|rank of next distal seq…

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|1|SS:1:3-3|1:8-8 AAXAAAAX|1|SS:1:3-3|1:8-8

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|1|SS:1:3-3|1:8-8 AAXAAAAX|1|SS:1:3-3|1:8-8

AAXAAAAX|1|SS:1:3-3|1:8-8 AAXAAAAX|1|SS:1:3-8|1:8-8

AAXAAAAX|1|SS:1:3-3|1:8-3 AAXAAAAX|1|SS:1:3-8|1:8-3

Iteration2: RE-RANKING

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|3|SS:3:3-3|3:8-8 AAXAAAAX|3|SS:3:3-3|3:8-8

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|3|SS:3:3-3|3:8-8 AAXAAAAX|3|SS:3:3-3|3:8-8

AAXAAAAX|3|SS:2:3-3|5:8-8 AAXAAAAX|5|SS:4:3-8|3:8-8

AAXAAAAX|2|SS:3:3-3|4:8-3 AAXAAAAX|4|SS:2:3-8|5:8-3

Iteration3: RE-RANKING

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|4|SS:4:3-3|4:8-8 AAXAAAAX|4|SS:4:3-3|4:8-8

AAXAAAAX|1|SS:1:3-3;8-8 AAXAAAAX|4|SS:4:3-3|4:8-8 AAXAAAAX|4|SS:4:3-3|4:8-8

AAXAAAAX|3|SS:2:3-3|6:8-8 AAXAAAAX|6|SS:5:3-8|3:8-8

AAXAAAAX|2|SS:3:3-3|5:8-3 AAXAAAAX|5|SS:2:3-8|6:8-3