CLAIMS

What is claimed is:

1. A method for identifying glycopeptides in a sample, the method comprising:

converting a mass spectrum of MS1 precursors of the sample into a plurality of nodes in a graph, each node corresponding to one mass and one retention time of a glycopeptide to be identified in the sample;

calculating differences in the mass ~~and/or~~and calculating differences in the retention time between all combinations of pairs of the nodes in the graph;

generating a graph theoretic network of the nodes wherein the generated graph theoretic network comprises graph theoretic edges corresponding to the calculated differences; and

predicting compositions of the glycopeptides in the sample based on the graph theoretic network of the nodes so as to identify the glycopeptides.

2. The method according to claim 1, further comprising:

for each node of the graph, setting the node as a hub node and all other nodes as arm nodes, and connecting the hub node with the arm nodes to form the pairs of the nodes.

3. The method according to claim 1 or 2, wherein the step of generating the graph theoretic network of the nodes comprises:

retaining a plurality of the pairs of the nodes, wherein the difference in the mass between each of the retained plurality of pairs of the nodes is equal to a mass of a glycan attachment within a list of known glycan attachments, and/or the difference in the retention time between each of the retained plurality of pairs of the nodes is less than a retention time threshold.

4. The method according to claim 3, wherein the list of known glycan attachments comprises one or more of: N-acetylhexosamines, sialic acids, Hexoses, and Deoxyhexoses.

5. The method according to claim 4, wherein the N-acetylhexosamines include N-acetylglucosamine and N-acetylgalactosamine, the sialic acids include N-acetylneuraminic acid and N-glycolylneuraminic acid, the Hexoses include Mannose and Galactose; and the Deoxyhexoses include Fucose.

6. The method according to any one of claims 3 to 5, wherein the retention time threshold is 50 seconds.

7. The method according to any one of claims 3 to 6, wherein the retention time threshold is variable based on a separation performance of a liquid chromatography device that is used prior to obtaining the MS1 precursors of the sample.

8. The method according to any one of claim 3 to 7, further comprising:

extracting the retained plurality of the pairs of the nodes as one or more subgraphs of nodes, each subgraph of nodes being isolated to other subgraphs of nodes,

wherein each subgraph of nodes represents a group of glycopeptides sharing a same peptide backbone with different glycan attachments, and

wherein the graph theoretic network of the nodes comprises the one or more subgraphs of nodes.

9. The method according to claim 8, wherein the step of predicting the compositions of the glycopeptides comprises:

identifying a sequence of the peptide backbone in each subgraph of nodes;

calculating a mass of the peptide backbone based on the identified sequence; and

predicting the compositions of the glycopeptides by identifying the glycan attachments of the glycopeptides based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph.

10. The method according to claim 8 or 9, further comprising:

for each subgraph of nodes, selecting a reference node;

identifying a sequence of the reference node of the subgraph; and

predicting compositions of the rest of the nodes in the subgraph based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph and the sequence of the reference node.

11. The method according to claim 10, wherein the step of predicting compositions of the rest of the nodes in the subgraph comprises:

predicting relative compositions of the rest of the nodes in the subgraph based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph; and

predicting absolute compositions of the rest of the nodes in the subgraph by incorporating the relative compositions of the rest of the nodes with the sequence of the reference node.

12. The method according to any one claims 1 to 11, further comprising:

providing the predicted compositions of the glycopeptides to a database including compositions of known glycans and peptides; and

conducting one or more searches to identify glycopeptides in the sample based on the database.

13. An apparatus for identifying glycopeptides in a sample, the apparatus comprising:

at least one input module;

at least one output module;

at least one processor; and

at least one memory including computer program code;

wherein the input module is configured to receive data from a liquid chromatography–mass spectrometry (LC-MS) system, the data including mass spectrum data,

wherein the output module is configured to output a result of the identified glycopeptides, and

wherein the at least one memory and the computer program code are configured to, with the at least one processor, cause the apparatus at least to:

convert a mass spectrum of MS1 precursors of the sample into a plurality of nodes in a graph, each node corresponding to one mass and one retention time of a glycopeptide to be identified in the sample;

calculate differences in the mass ~~and/or~~and calculating differences in the retention time between all combinations of pairs of the nodes in the graph;

generate a graph theoretic network of the nodes wherein the generated graph theoretic network comprises graph theoretic edges corresponding to the calculated differences; and

predict compositions of the glycopeptides in the sample based on the graph theoretic network of the nodes so as to identify the glycopeptides.

14. The apparatus according to claim 13, wherein the apparatus is further caused to:

for each node of the graph, set the node as a hub node and all other nodes as arm nodes, and connect the hub node with the arm nodes to form the pairs of the nodes.

15. The apparatus according to claim 13 or 14, wherein at the step of generating the graph theoretic network of the nodes, the apparatus is caused to:

retain a plurality of the pairs of the nodes, wherein the difference in the mass between each of the retained plurality of pairs of the nodes is equal to a mass of a glycan attachment within a list of known glycan attachments, and/or the difference in the retention time between each of the retained plurality of pairs of the nodes is less than a retention time threshold.

16. The apparatus according to claim 15, wherein the list of known glycan attachments comprises one or more of: N-acetylhexosamines, sialic acids, Hexoses, and Deoxyhexoses.

17. The apparatus according to claim 16, wherein the N-acetylhexosamines include N-acetylglucosamine and N-acetylgalactosamine, the sialic acids include N-acetylneuraminic acid and N-glycolylneuraminic acid, the Hexoses include Mannose and Galactose; and the Deoxyhexoses include Fucose.

18. The apparatus according to any one of claims 15 to 17, wherein the retention time threshold is 50 seconds.

19. The apparatus according to any one of claims 15 to 18, wherein the retention time threshold is variable based on a separation performance of a liquid chromatography device that is used prior to obtaining the MS1 precursors of the sample.

20. The apparatus according to any one of claims 15 to 19, wherein the apparatus is further caused to:

extract the retained plurality of the pairs of the nodes as one or more subgraphs of nodes, each subgraph of nodes being isolated to other subgraphs of nodes,

wherein each subgraph of nodes represents a group of glycopeptides sharing a same peptide backbone with different glycan attachments, and

wherein the graph theoretic network of the nodes comprises the one or more subgraphs of nodes.

21. The apparatus according to claim 20, wherein at the step of predicting the compositions of the glycopeptides, the apparatus is caused to:

identify a sequence of the peptide backbone in each subgraph of nodes;

calculate a mass of the peptide backbone based on the identified sequence; and

predict the compositions of the glycopeptides by identifying the glycan attachments of the glycopeptides based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph.

22. The apparatus according to claim 20 or 21, wherein the apparatus is further caused to:

for each subgraph of nodes, select a reference node;

identify a sequence of the reference node of the subgraph; and

predict compositions of the rest of the nodes in the subgraph based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph and the sequence of the reference node.

23. The apparatus according to claim 18, wherein at the step of predicting compositions of the rest of the nodes in the subgraph, the apparatus is caused to:

predict relative compositions of the rest of the nodes in the subgraph based on the difference in the mass and/or retention time between each pair of the nodes in the subgraph; and

predict absolute compositions of the rest of the nodes in the subgraph by incorporating the relative compositions of the rest of the nodes with the sequence of the reference node.

24. The apparatus according to any one of claims 13 to 23, wherein the apparatus is further caused to:

provide the predicted compositions of the glycopeptides to a database including compositions of known glycans and peptides; and

conduct one or more searches to identify glycopeptides in the sample based on the database.

25. A computer readable medium comprising instructions which, when executed by a processor, make the processor perform a method for identifying glycopeptides in a sample according to any one of claims 1 to 12.