

De novo glycan structure search with CID MS/MS spectra of native N-glycopeptides

Hannu Peltoniemi^{1,2}, Sakari Joenväärä^{1,3} and Risto Renkonen¹

Haartman Institute, University of Helsinki¹, Applied Numerics Oy², MediceL Oy³, Helsinki, Finland

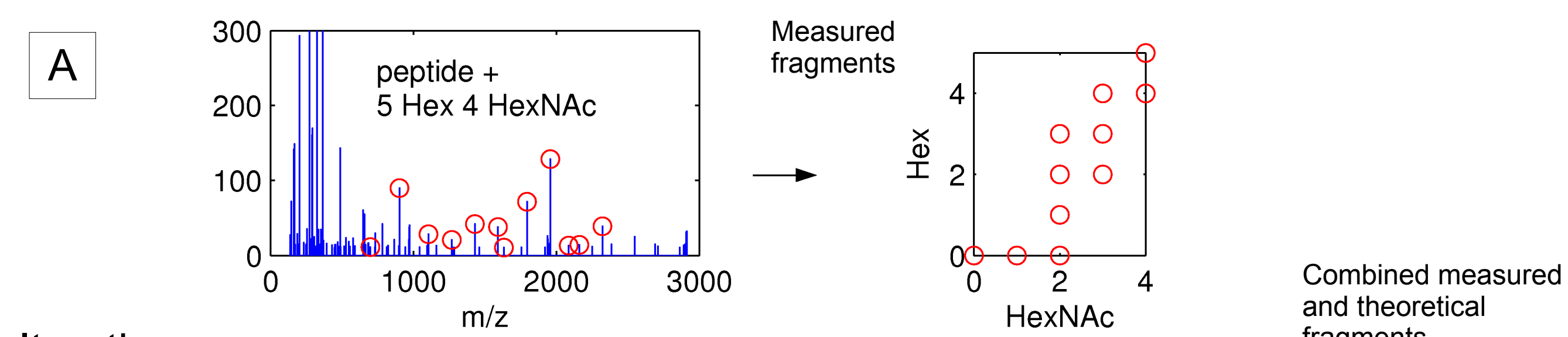
Introduction

A novel branch-and-bound type algorithm to search *de novo* N-glycan structures (without linkage information) is developed. The goal is to analyze intact N-glycopeptides without the need to release the glycan moiety component from the peptide. The algorithm is based on iterative growth and selection of a population of glycan structures. The algorithm does not use a database of known structures, only the N-glycan core structure is given as input. The method was tested with a numerically generated spectra, with a purified human sero-transferring sample and with complex mixture of N-glycopeptides enriched from human plasma.

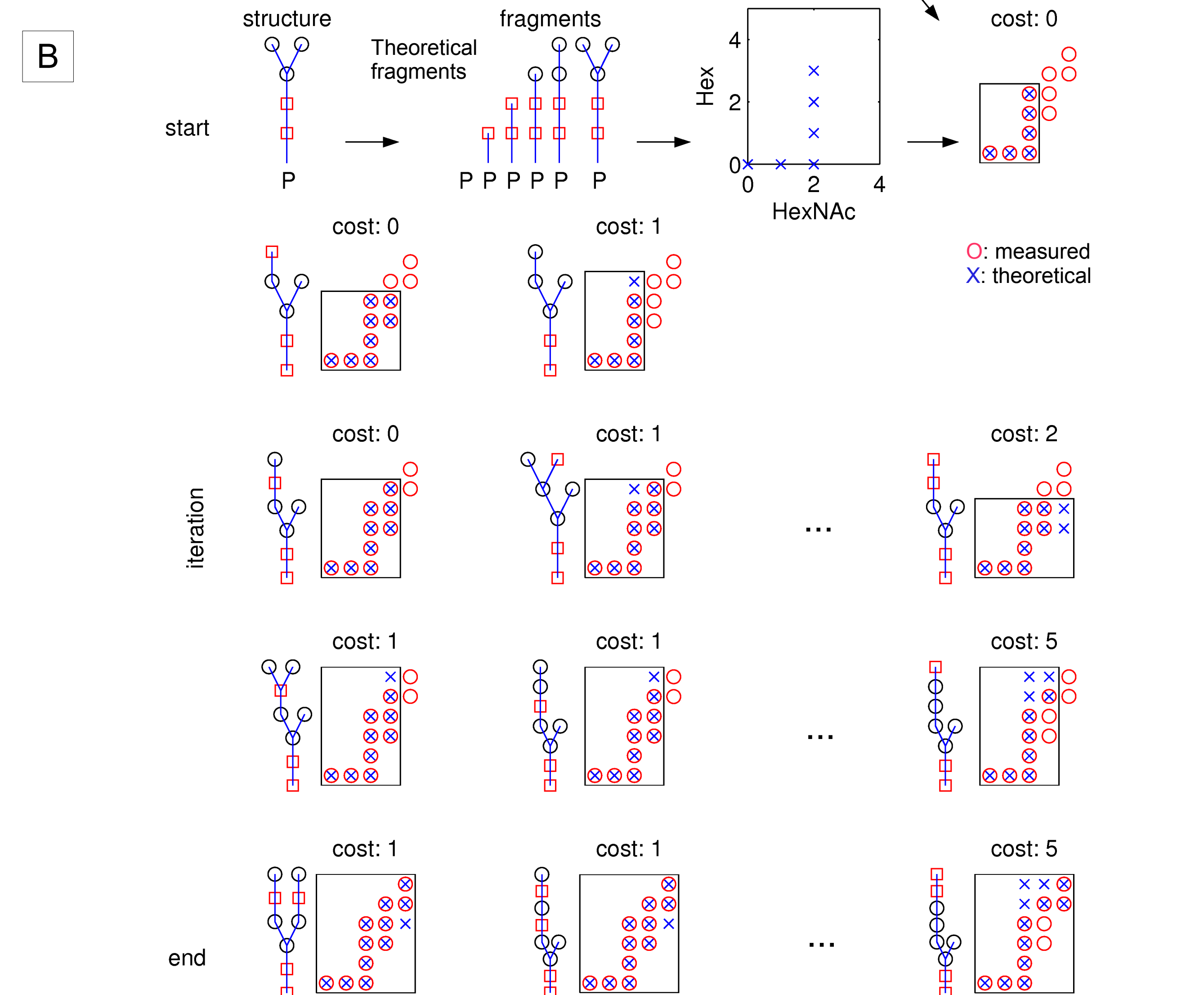
Methods

The input data is MaxEnt3 deconvoluted MS/MS spectra of protonated N-glycopeptides. Prior to structure search the peptide sequences and potential glycan compositions with matching fragments are evaluated. The algorithm grows a population of N-glycan structures, starting with a given N-glycan core structure, attaching iteratively monosaccharides until the structures have grown to target composition (Fig 1). The theoretical spectra are generated using fragments resulting from a unlimited number of glycosidic cleavages. After the structure growth the glycopeptide population is ordered by score values. The score is defined as a negative logarithm of a probability that a random set of fragments would have as many or more shared peaks with measured spectrum as the ranked glycopeptide. The algorithm options include the number of possible connections for a given monosaccharide and the maximum number of branches for final structures.

Initialization



Iteration



Scoring

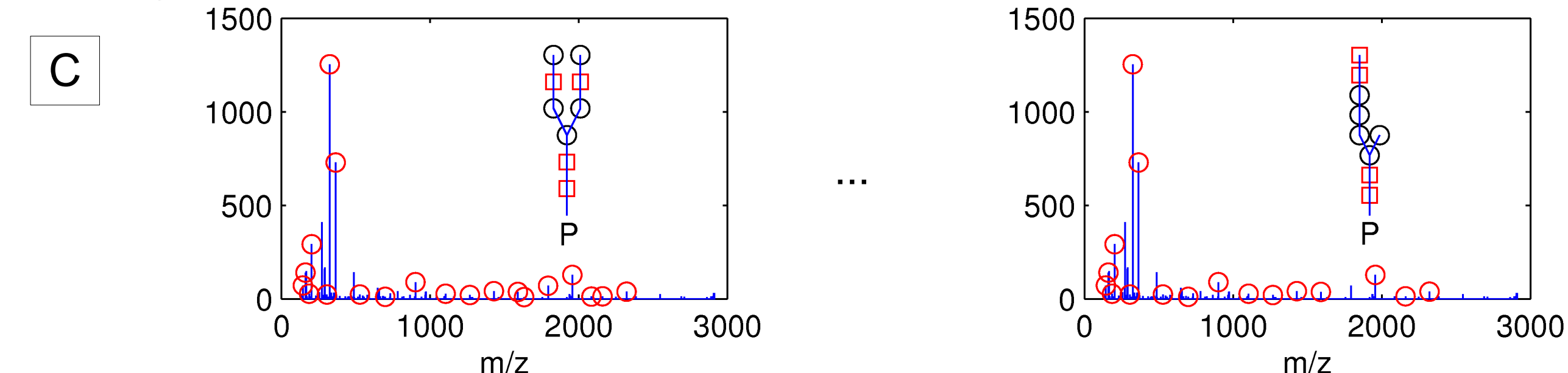


Figure 1. De novo glycan structure search.

A) Initialization

Peaks matching intact peptide + glycan composition fragments are selected from the spectrum.

B) Iteration

Structure search is started from a N-glycan core. At each iteration round one new carbohydrate is added. If the number of structures grows above a given limit the structures with highest cost are removed. The cost is defined as the number of not matching theoretical and measured reducing end fragments, i.e. high cost is a sign of a low match between measured and theoretical spectra.

C) Scoring

When the target composition is reached the structures are ordered by score. All fragments, also non reducing end fragments, are included in the score calculation.

Results

The method sensitivity were tested with numerically generated spectra (Fig 2). The results are good if there are more than about one half of the reducing end fragments present, especially without added noise fragments.

For the analyzed sero-transferring and plasma samples we were able to reproduce structures known in the literature. However, quite often the correct structure were not the first one but were among the highest scoring ones. High quality spectra were essential for structure identification. Putative glycan structures were generated for 107 (76%) from the total 140 input precursors with identified peptide and glycan composition. An example result for a human plasma sample precursor is shown at the Fig 3.

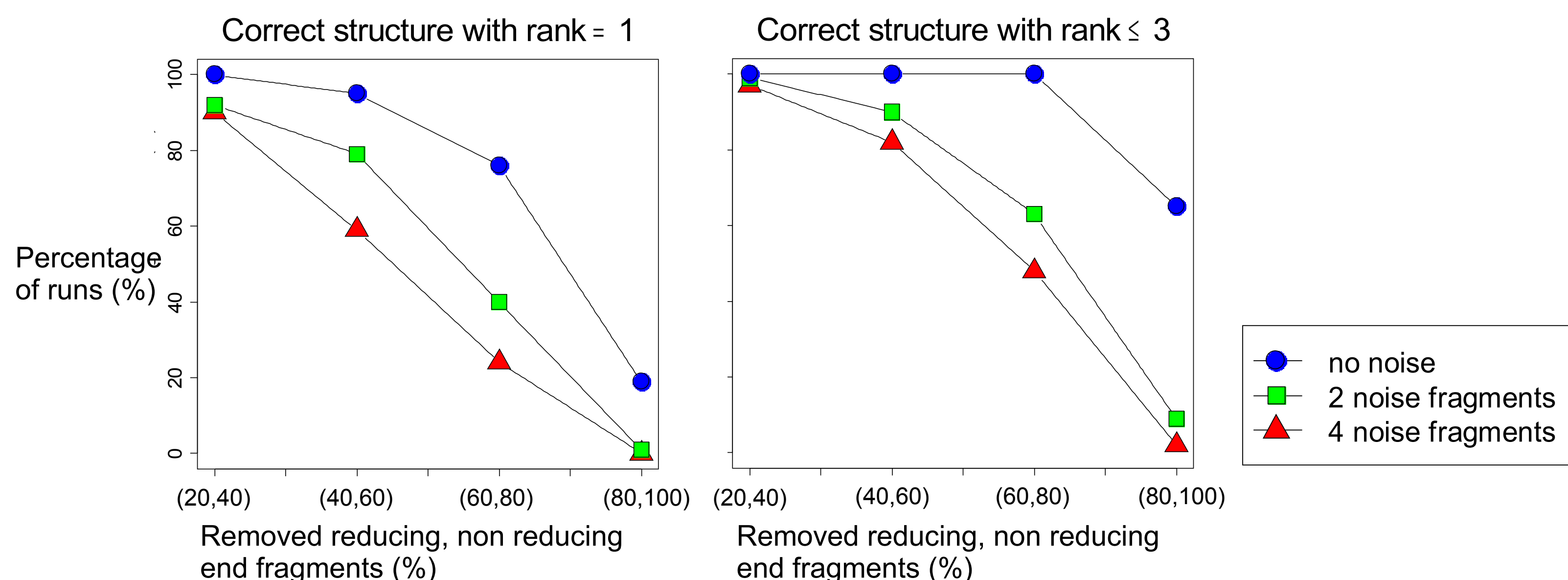


Figure 2. Amount of algorithm runs where the correct structure had highest score (left) or had at least third highest score (right). The input data is randomly distorted theoretical spectra of a glycan with three branches (Fig 3, structure 1). Each mark is a result of 100 algorithm runs.

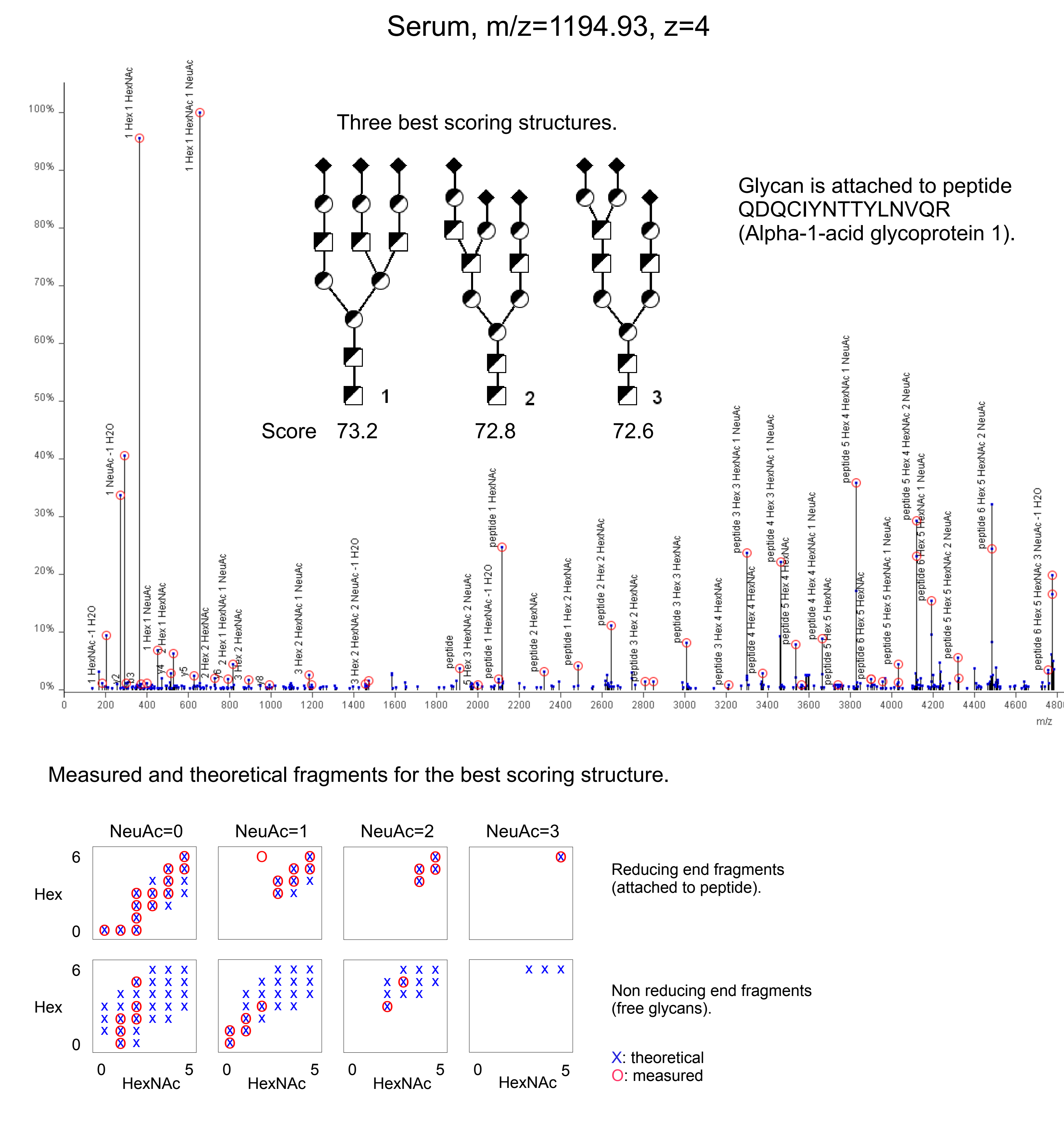


Figure 3. Spectrum and structure search result for a serum sample precursor.

Conclusions

The method was found to be able to find the correct glycan structure with high quality spectra. However, in practise the results typically contained several glycans matching almost equally well. Despite the potential lack of unique results the method can strongly reduce the number of possible glycan structures from the vast amounts of potential ones.

References

Joenväärä et al., N-Glycoproteomics – An automated workflow approach., *Glycobiology* 2008,18(4):339-349.

For further information, please contact: hannu.peltoniemi@helsinki.fi