

# **Factors that affect Protein Identification using MS/MS Peptide Mapping**

Haofei Wang, Scot R. Weinberger and Ron Orlando

Complex Carbohydrate Research Center,  
University of Georgia, Athens, GA

\*CIPHERGEN Biosystems, Inc., Fremont, CA

## Overview

As a continuing work to the analysis on factors that affect protein identification using MS spectrum (presented on ASMS'2000 as poster), here we obtain information on the effect of common variables affecting protein identification by MS/MS database search.

## Objective:

### **Investigate the dynamic interaction of**

- Peptide tolerance (Error window on experimental peptide mass values )
- MS/MS tolerance (Error window for MS/MS fragment ion mass values )
- Limit taxonomy on database search(es)
- Allowing the presence of post-translational modifications

### **on the success of database mining experiments**

## Introduction:

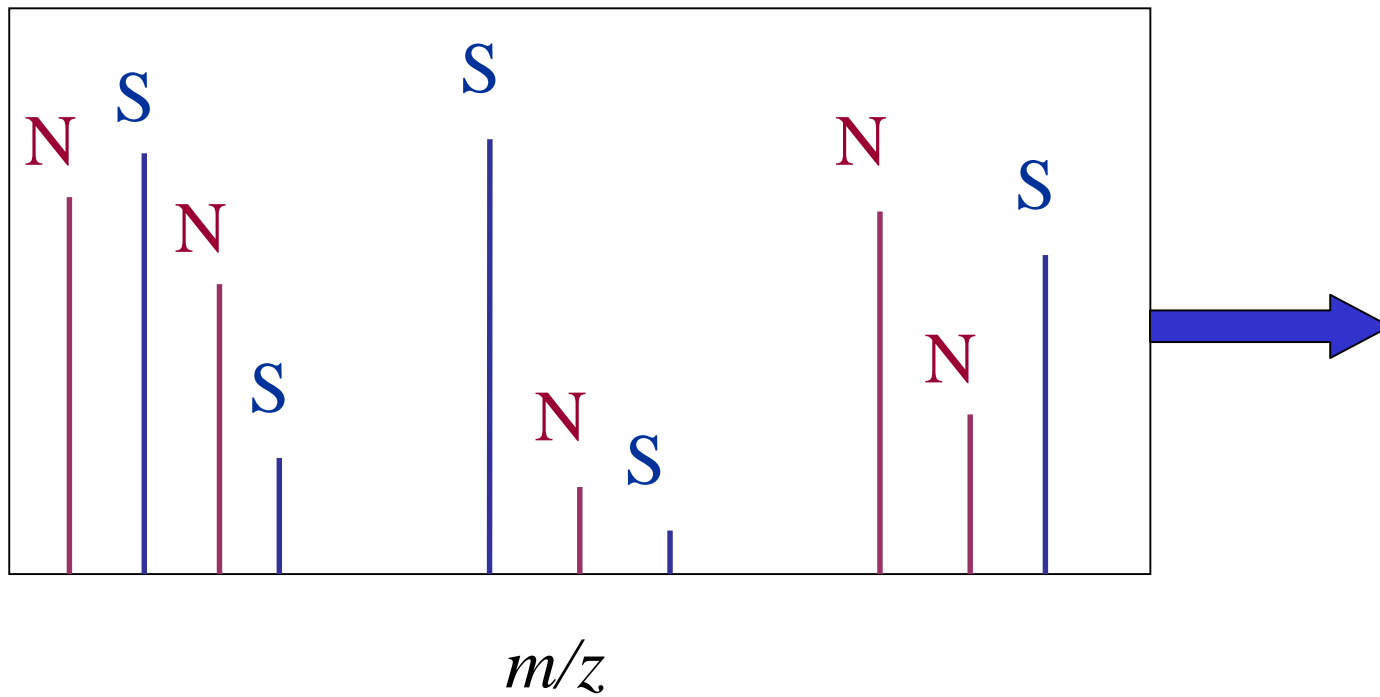
Mass spectrometry has been combined with database search utilities to create a valuable protein identification tool. Here, we discuss the variables that affect database searching with MS/MS spectra. Intact proteins are digested by trypsin into peptides. The peptide fragment masses are then obtained by MS/MS and searched against theoretical fragments from protein cDNA and EST database entries. Database mining success appears to depend upon protein purity, target organism and accuracy of peptide mass assignment and MS/MS fragment m/z assignment.

## Methods:

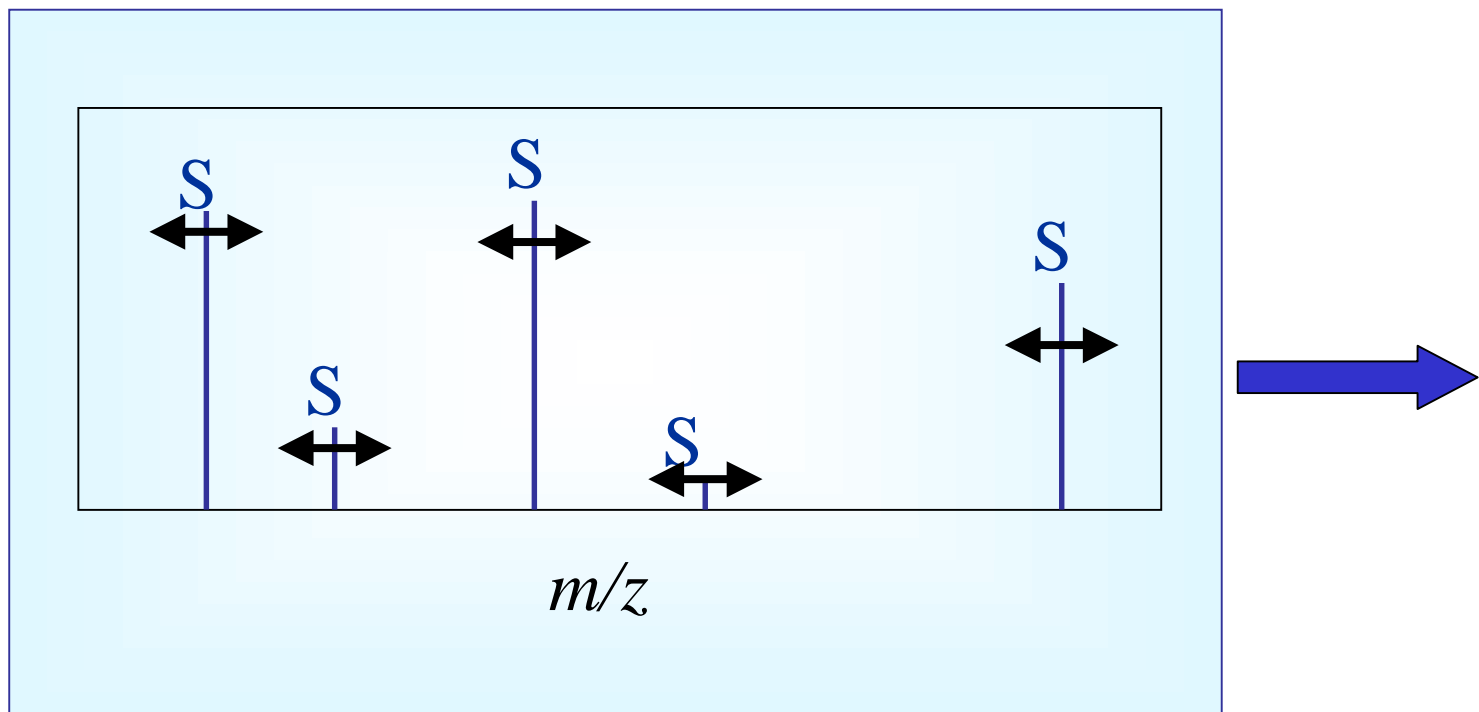
We have selected three proteins extracted from *Arabidopsis thaliana* leaves. Proteins were subjected to 2-D gel electrophoresis, in gel digestion, peptide extraction and LC-MS/MS. Obtained spectra were idealized by optimizing the mass assignments then degraded by randomizing the mass accuracy. These MS/MS spectra were then used for database searches to see the effect of the peptide tolerance, MS/MS tolerance, changing the number of peaks in the spectrum, limiting species placed on the search and the presence of post-translational modifications. Selected *Arabidopsis thaliana* proteins are:

Tested Protein	Protein ID	NCBI ID#	Taxonomy	Nominal mass (Mr)	Calculated pI value
<b>Protein A</b>	ATPase beta subunit	gi 7525040	<i>Arabidopsis thaliana</i>	53900	5.38
<b>Protein B</b>	sedoheptulose-bisphosphatase precursor	gi 7525041		42388	6.38
<b>Protein C</b>	23 kDa polypeptide of oxygen-evolving complex	gi 7525042		28078	7.38

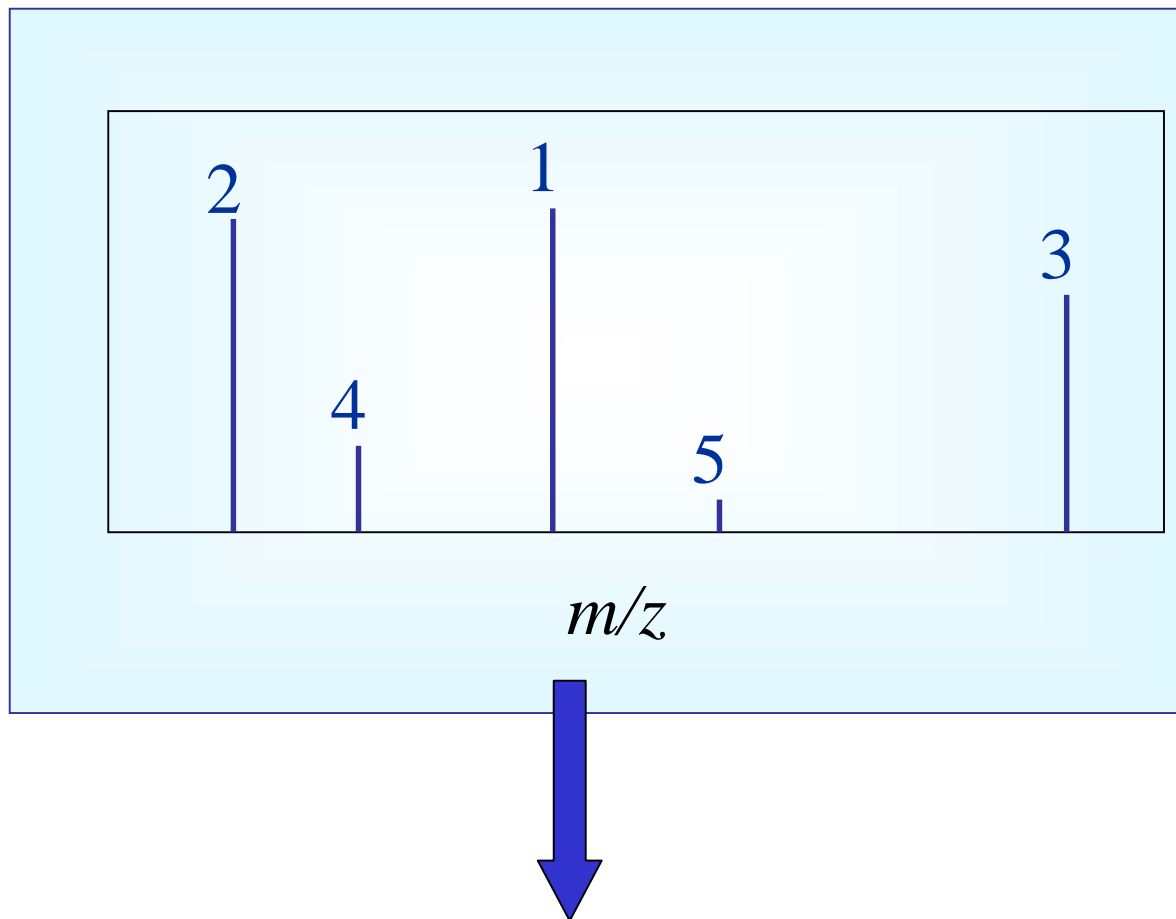
# 1. Real MS/MS Spectrum

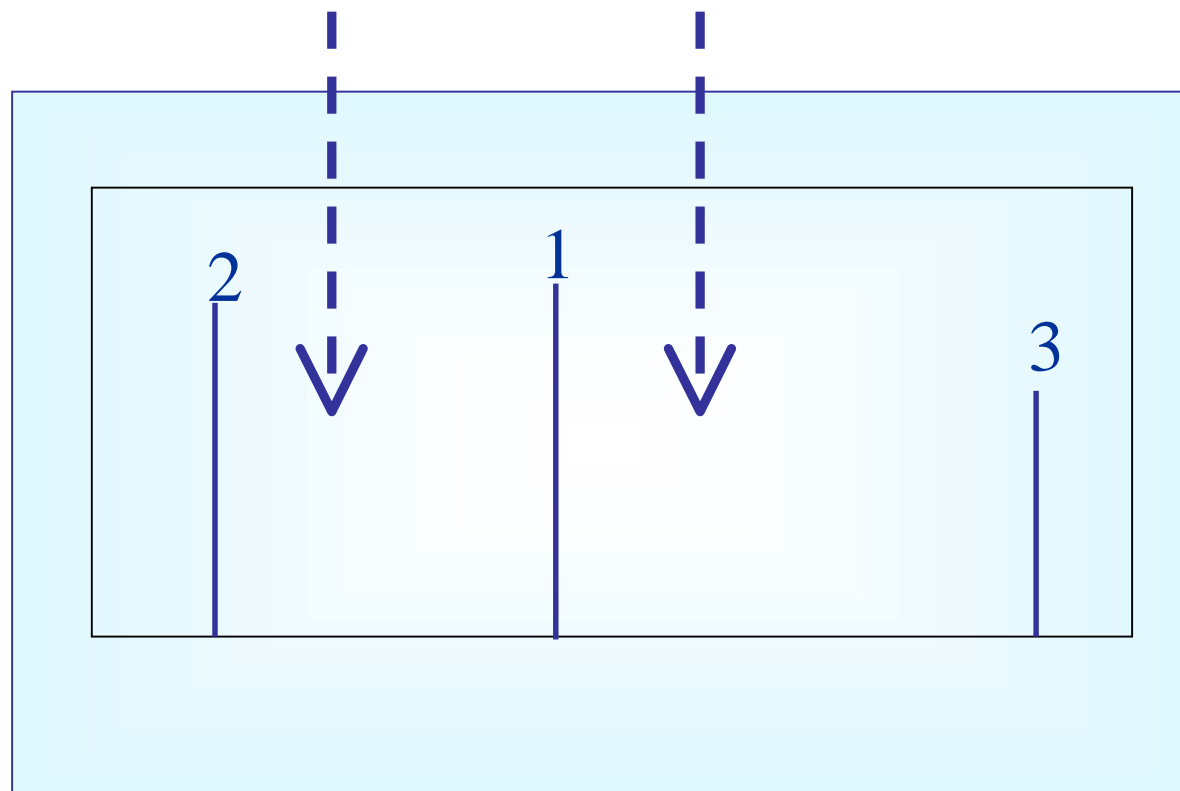


## 2. Remove noise and randomizing peptide mass accuracy and fragment $m/z$ accuracy

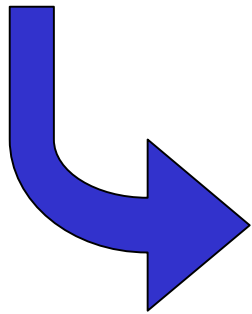


### 3. Rank fragment peaks by intensity





**4. Creating new MS/MS peak list  
containing specific # of fragments by  
deleting the peak(s) with low intensity**



**Search Real  
spectrum**

**Search idealized  
spectrum**

**5. Search NCBI nr Database using  
Mascot®**

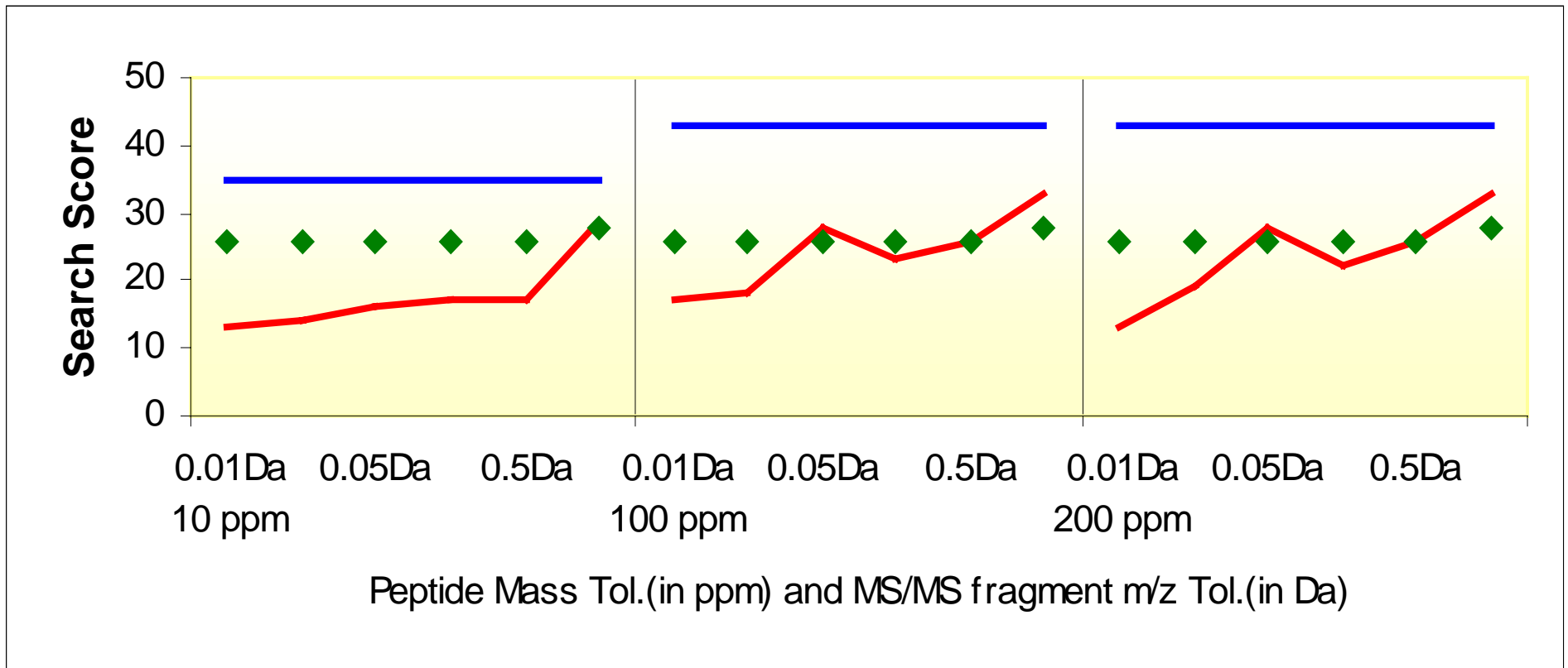
# Results

MS/MS spectrum or optimized peak lists are searched using **Mascot® MS/MS Ion Search** function. (<http://www.matrixscience.com>). *Search Score* is calculated as  $10 * \text{Log}(P)$ , where P is the probability that the observed match is a random event. Protein identification is successful when search score is higher or equals to the homology or identity threshold scores in each search.

# Effect of Peptide Mass Accuracy and Fragment Mass Accuracy

## Protein A

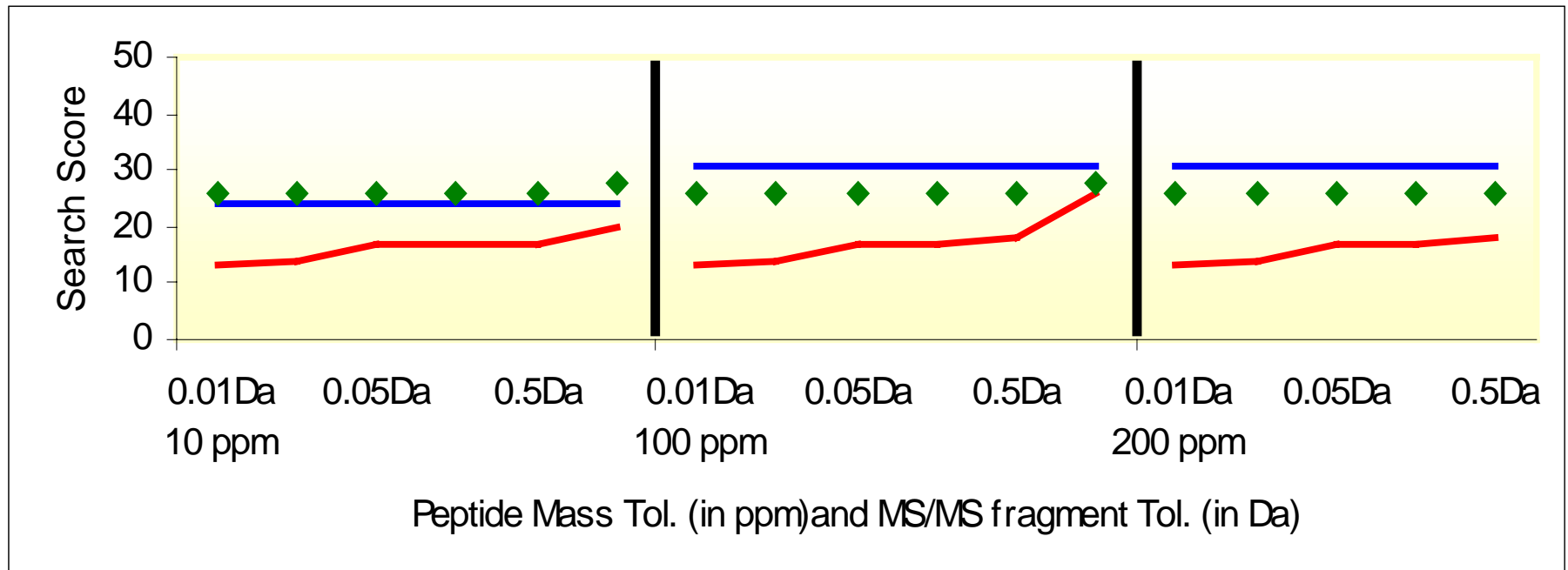
Search score vs. peptide mass Tol. and MS/MS fragment m/z Tol.



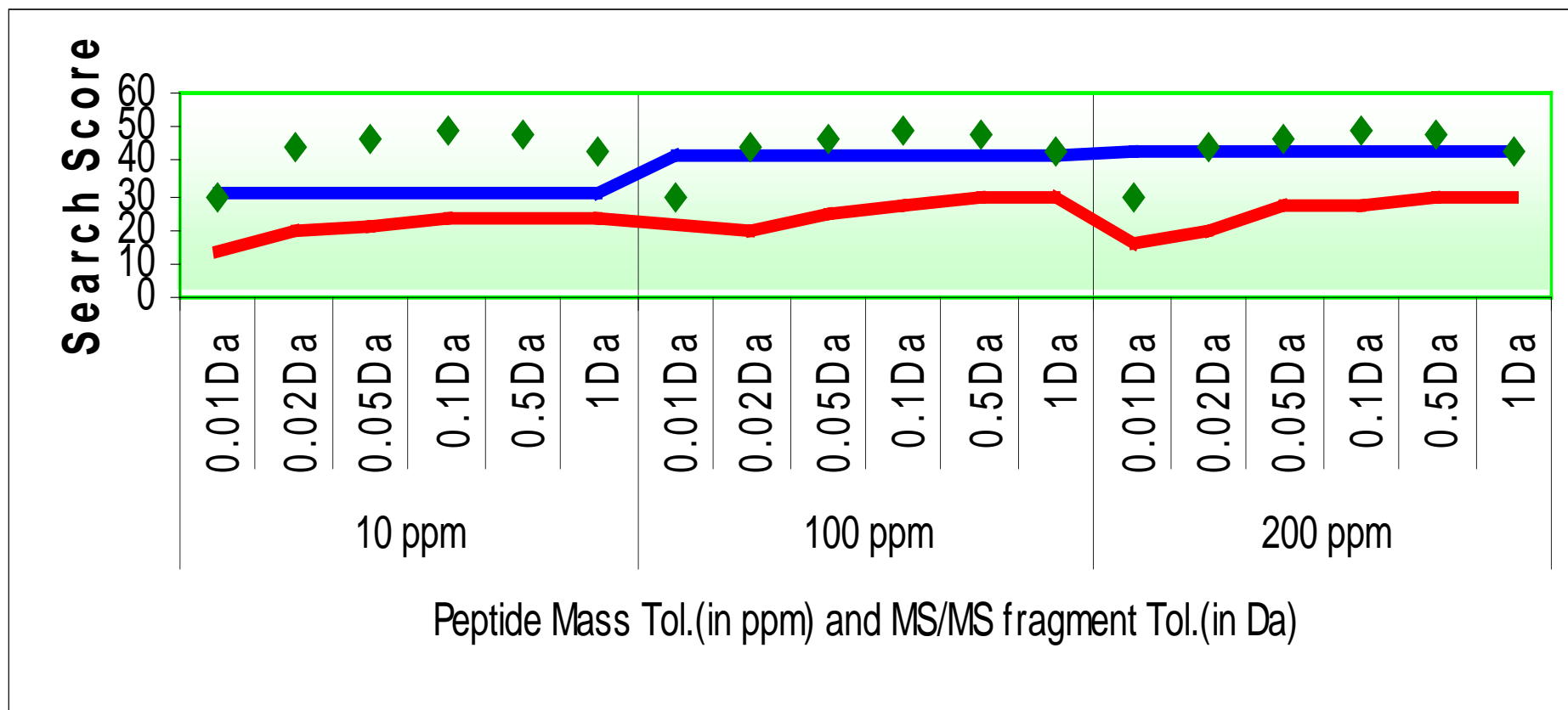
## Effect of limit Search to *Arabidopsis thaliana*

### Protein A

Search same spectrum with limit to *Arabidopsis thaliana*

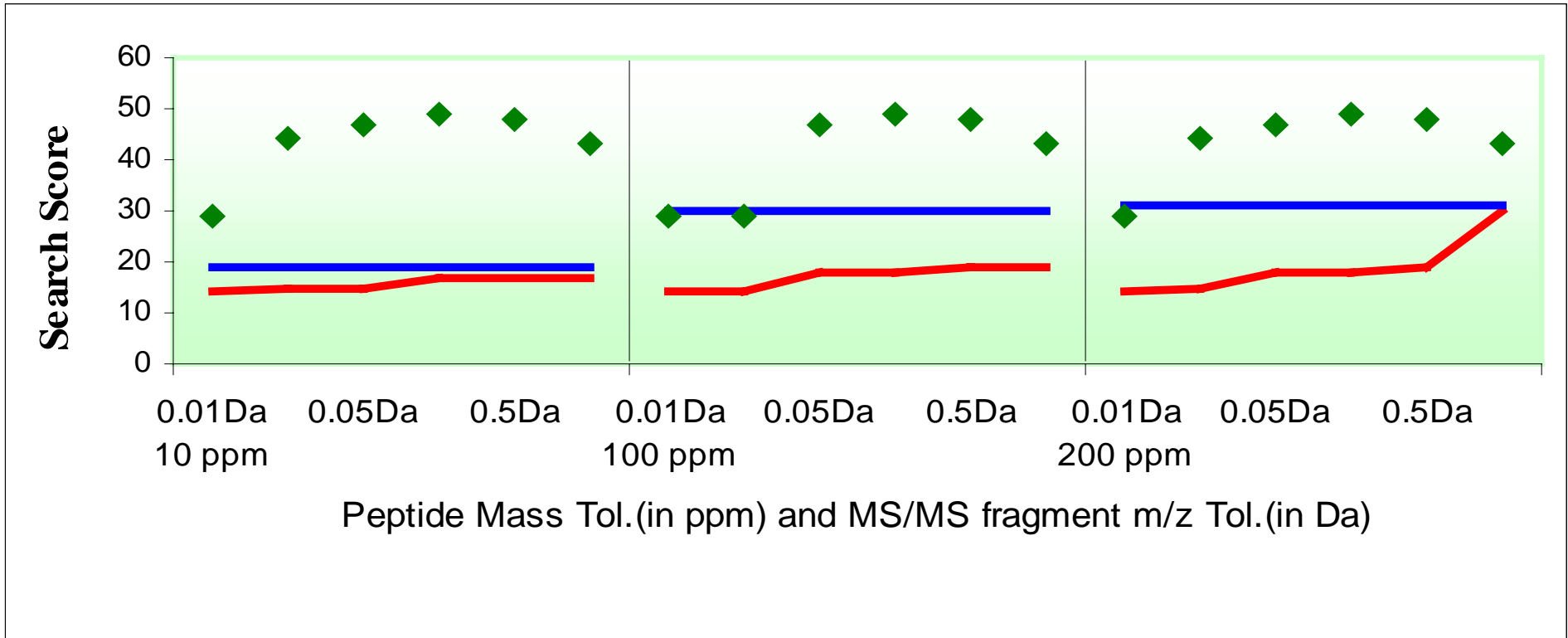


## Protein B



**Conclusion:** Relaxing precursor peptide mass accuracy from 10 to 100ppm raised the confidence requirement for identifying protein identity. Relaxing mass accuracy from 100 to 200 ppm shows no significant difference in searching result.

# Protein B

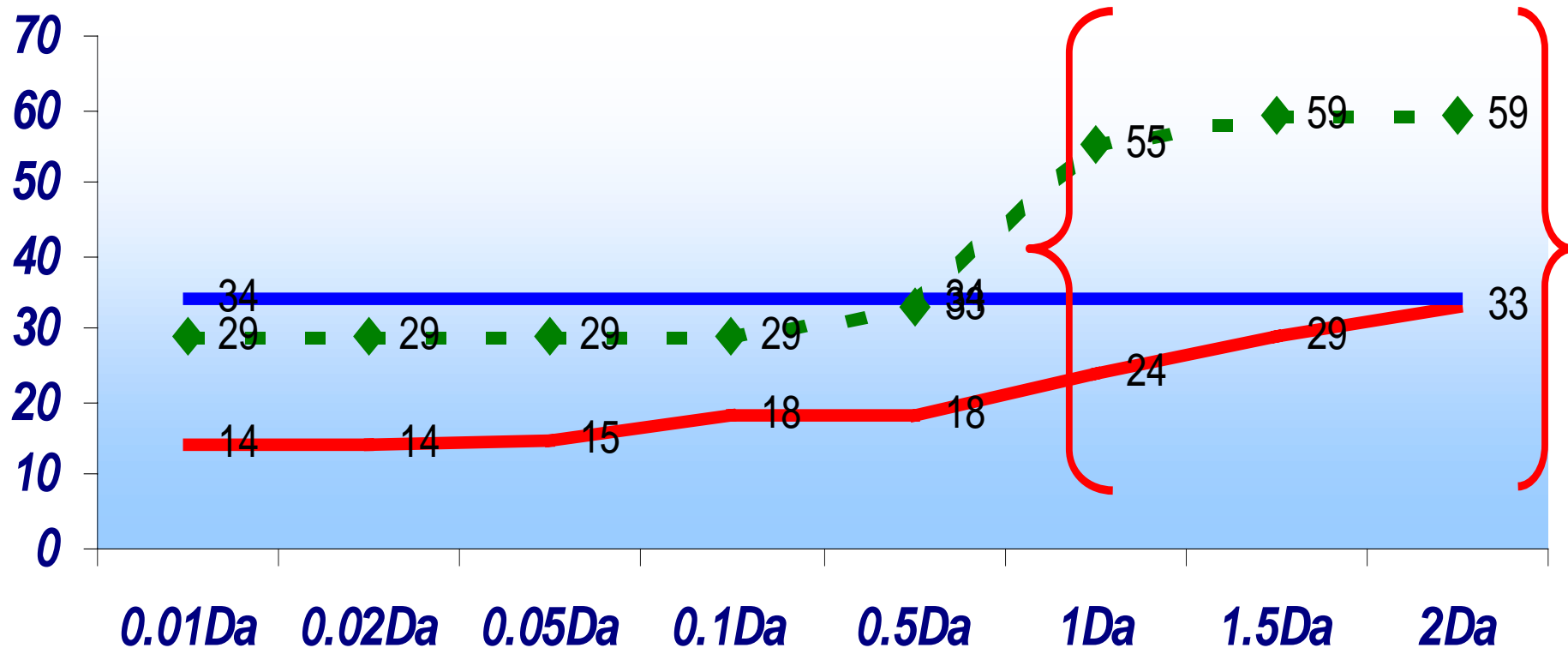


**Conclusion:** Limiting searches lowered both the homology confidence requirement and the identity confidence requirement, thus chance for identify a protein is improved

# Effect of MS/MS Tolerance

Protein C

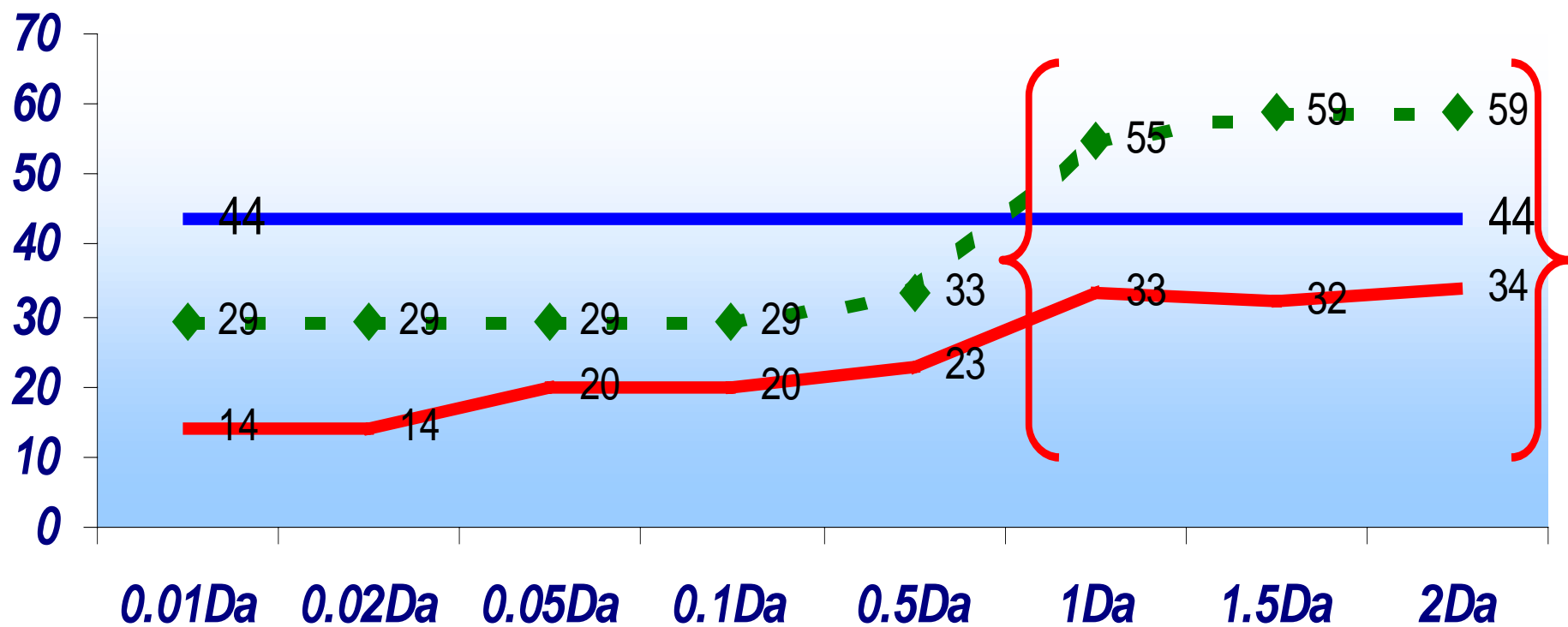
## Search Score vs. MS/MS Tol. (Peptide Tol. 10ppm)



# Effect of MS/MS Tolerance

Protein C

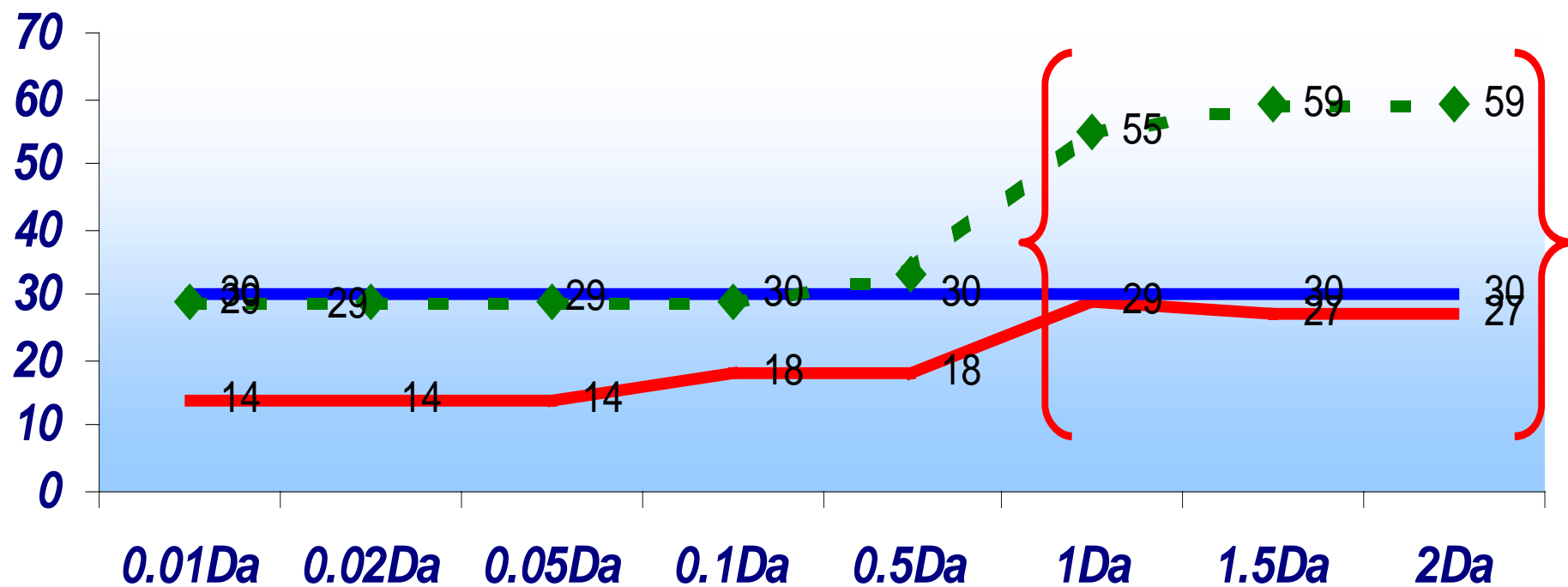
## Search Score vs. MSMS Tol. (Peptide Tol. 100ppm)



# Effect of MS/MS Tolerance

## Protein C

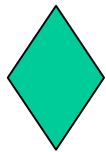
### Search Score vs. MS/MS Tol. (Peptide Tol.100ppm, limit to *Arabidopsis*)



**Conclusion:** Increasing MS/MS fragment m/z mass error window raises the homology and identity confidence threshold. Higher search score is observed when fragment m/z error window is wider than 1 Da, which is a result of including randomly matched peaks in response to the increase of MS/MS fragment m/z mass error window. Relaxing fragment m/z error window raises the chance for random match and weakens protein identification accuracy.

 Score threshold for  
significant **identity** ( $p < 0.05$ )

 Score threshold for  
significant **homology** ( $p < 0.05$ )



**Search Score**

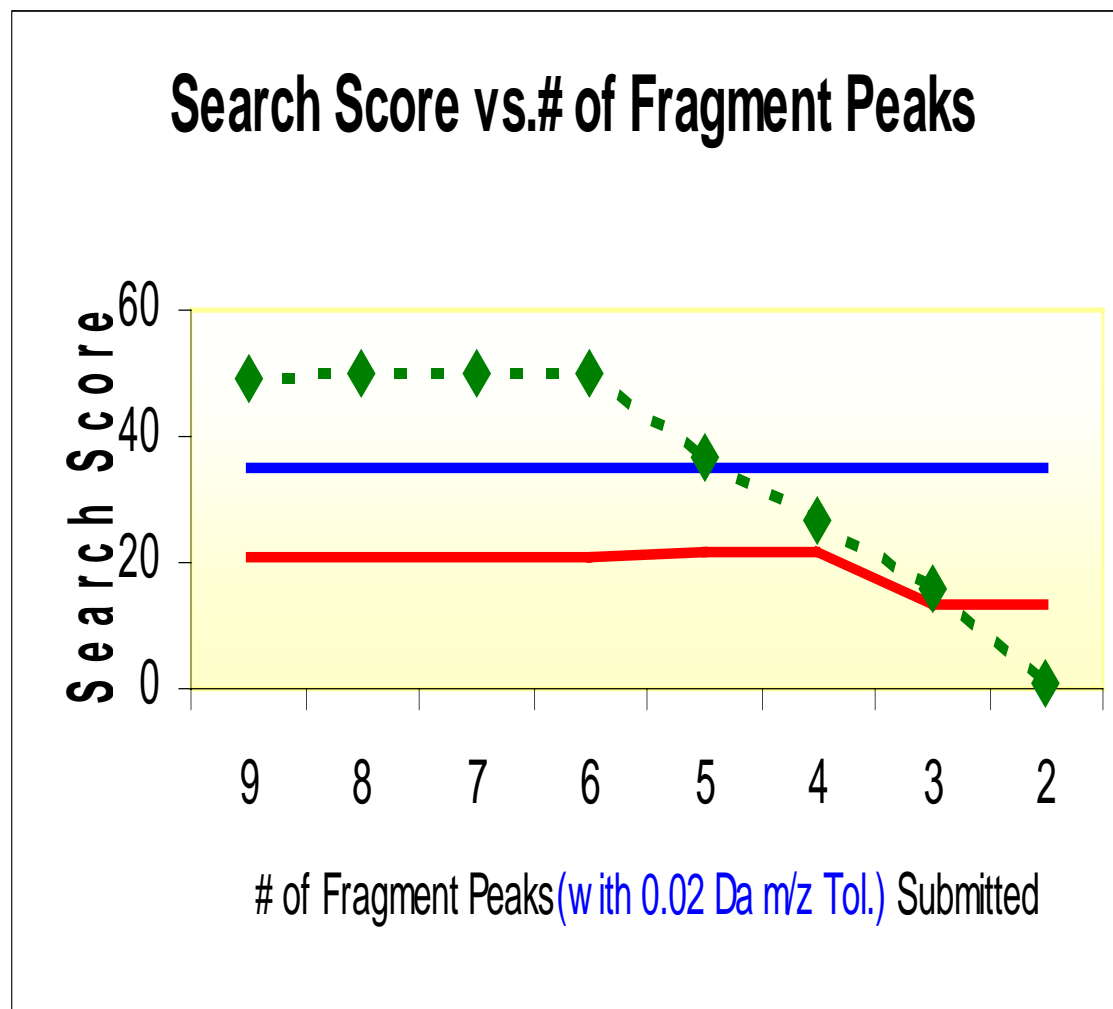
# Effect of the number of peaks in MS/MS spectrum

## Protein A

MS/MS Tol. (Da)	Min. # of MS/MS peaks to identify protein	
	Homology	Identity
0.02	3	5
0.2	3	6
2	5	> 9

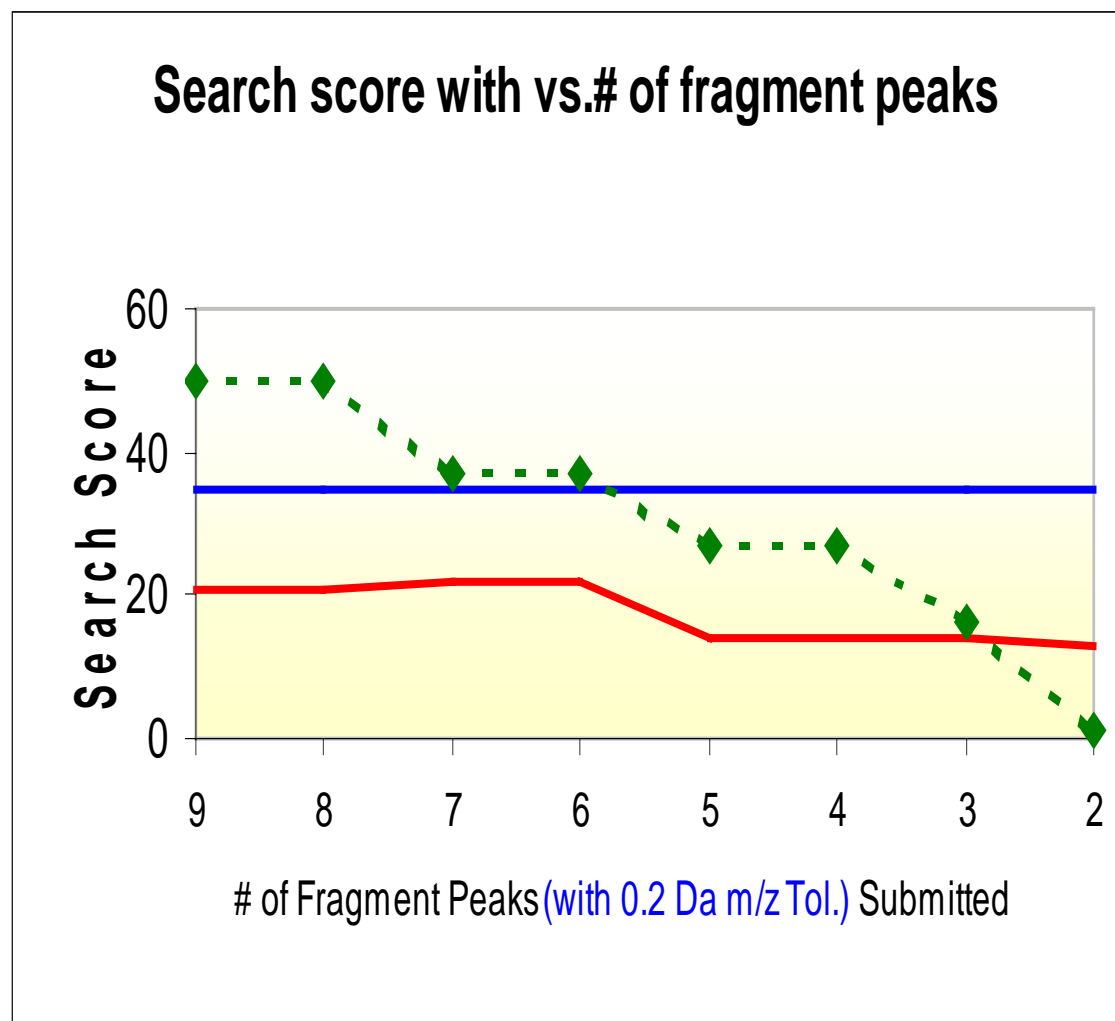
# Effect of the number of peaks in MS/MS spectrum

## Protein A



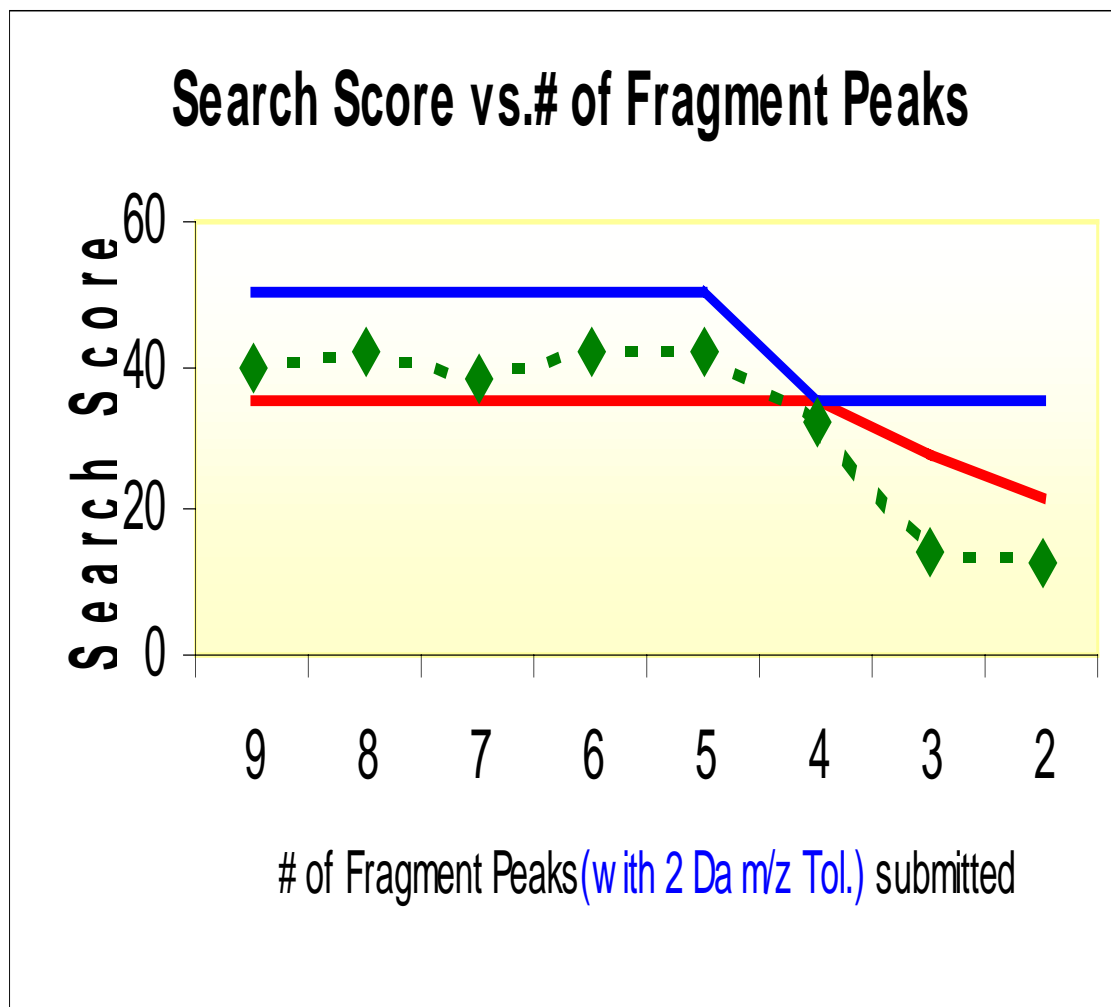
# Effect of the number of peaks in MS/MS spectrum

## Protein A



# Effect of the number of peaks in MS/MS spectrum

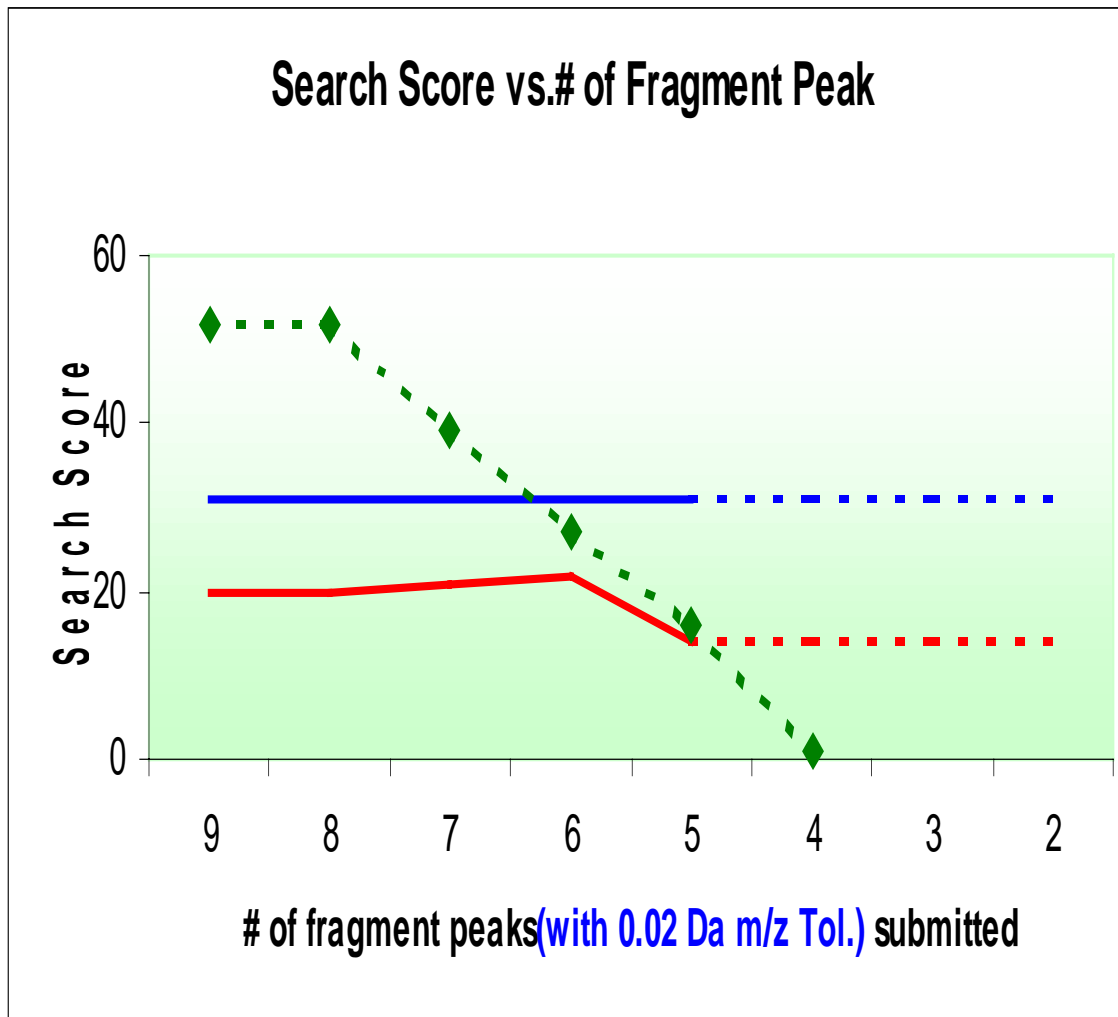
## Protein A



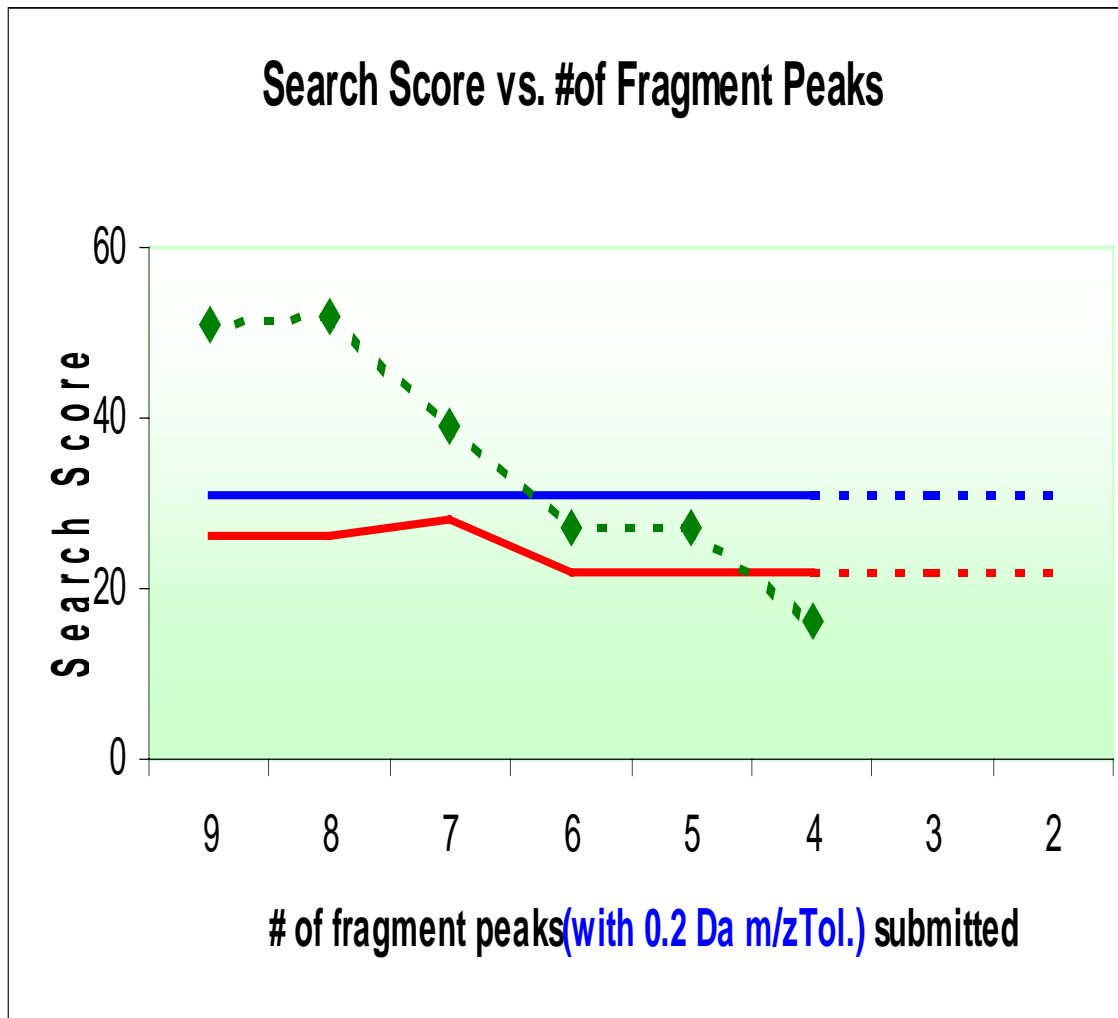
## Protein B

MS/MS Tol. (Da)	Min. # of MS/MS peaks to identify protein	
	Homology	Identity
0.02	5	7
0.2	5	7
2	7	7

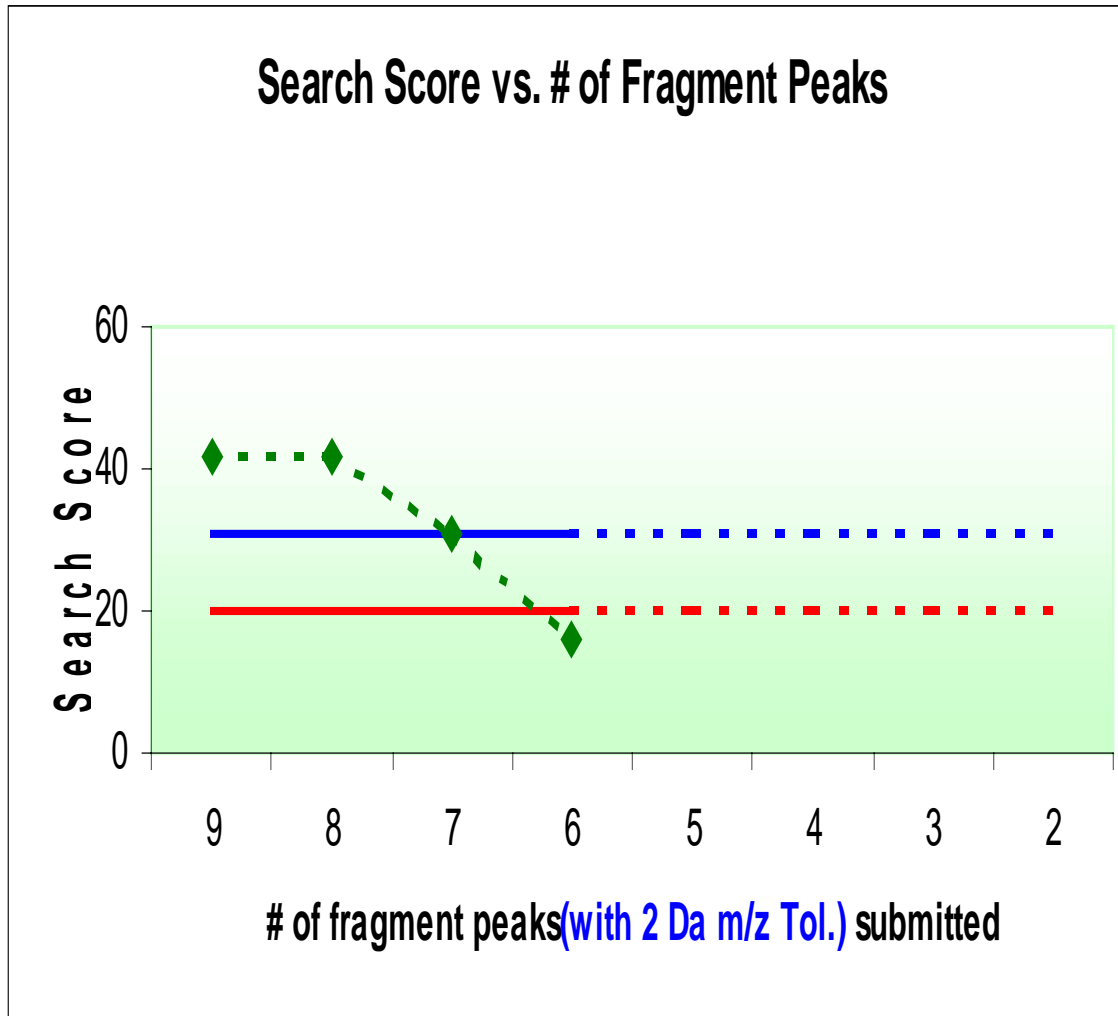
# Protein B



# Protein B



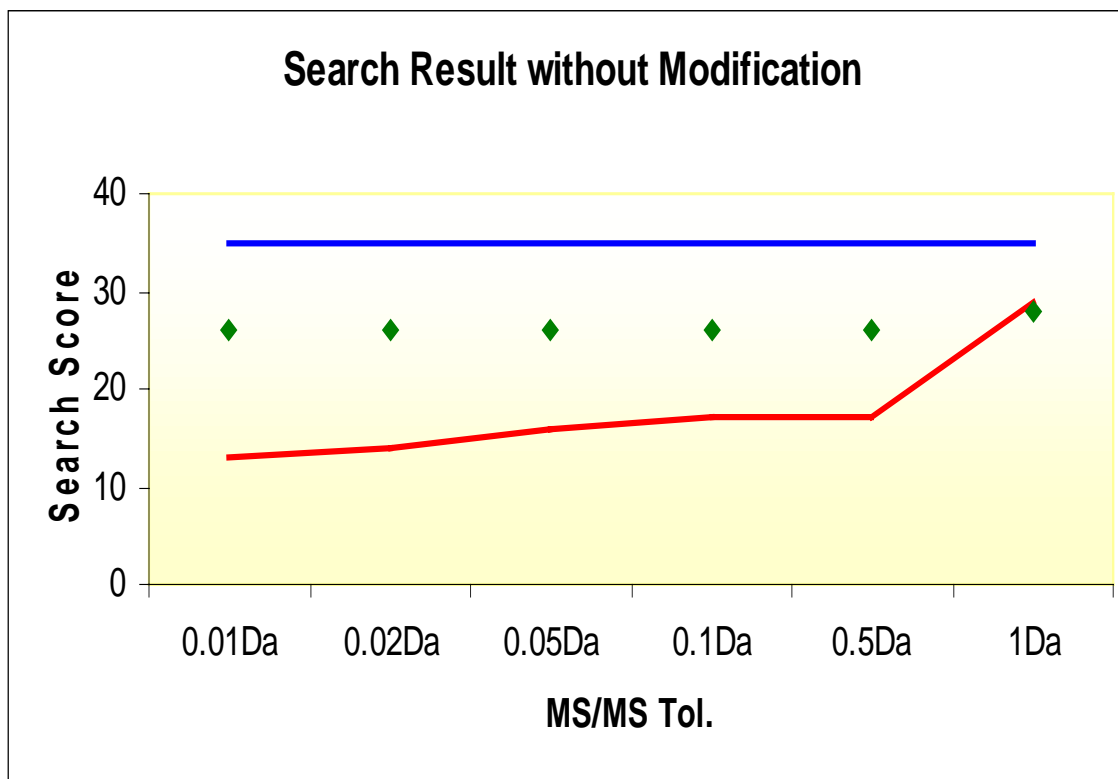
# Protein B



**Conclusion:** Search score is determined by the number of peaks in the submitted MS/MS spectrum in general. Other variables affect the confidence score thresholds for protein identification. Increasing number of peaks raises the search score, thus chance to meet homology and identity confidence requirement is increased. At least 3 and 5 peaks (with fragment  $m/z$  mass error window no wider than 0.02 Da) are required to identify protein A and B respectively.

# Effect of allowing post-translational modification

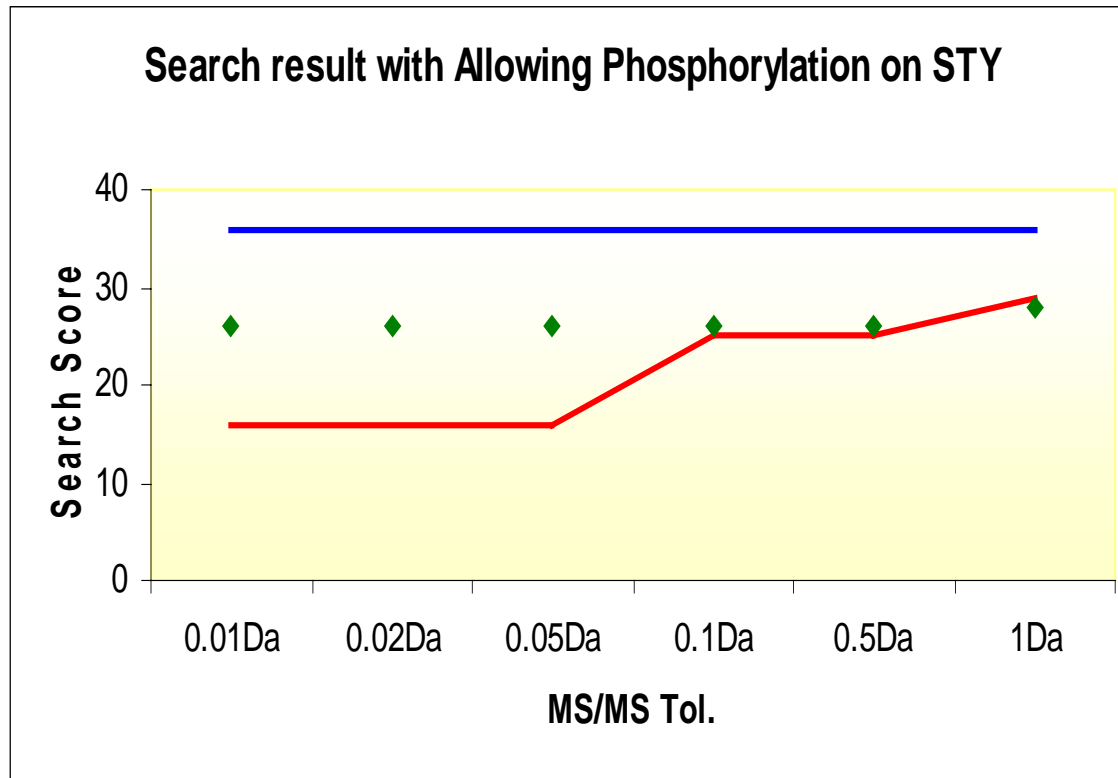
## Protein A



**No modification**

# Effect of allowing post-translational modification

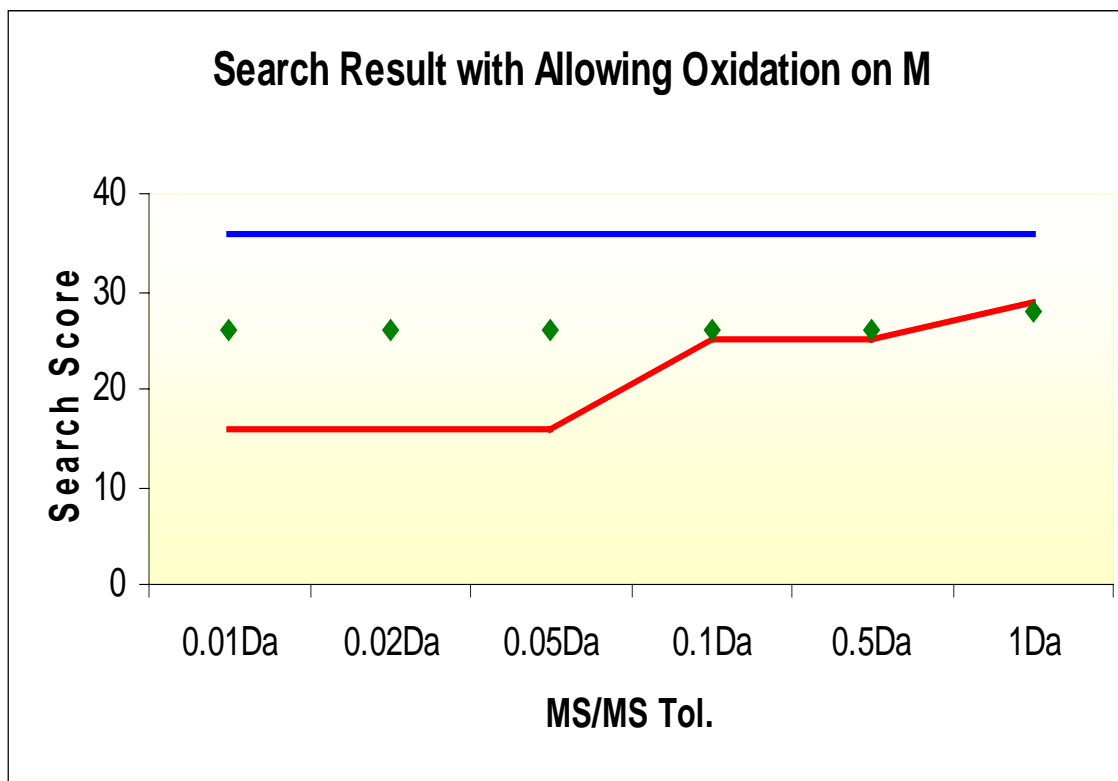
## Protein A



## Phosphorylation on STY

# Effect of allowing post-translational modification

## Protein A



## Oxidation on M

## **Conclusion:**

Allowing post-translational modification has no effect on searching score. It affects searching results by increasing the confidence requirement to identify protein homology, thus chance for identifying a protein is lowered.