

Getting to grips with histone modifications.

Richard Jacob

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

Getting to grips with histone modifications.

There has been an increase in the number of questions about analyzing Histone modifications.

- **Interest in Histone modifications and the Histone code.**
- **Robust sample preparation methods have been developed.**
 - Derivatization using propionic anhydride
 - Improvements in MS sensitivity and accuracy have helped in the analysis of the proteins and their post translation modifications.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

In the last couple of years we have seen an increase in the number of questions to our support email addresses about analyzing histone samples and their modifications.

Interest in Histone modifications and the so-called histone code is the driving force behind these queries. Another part of the reason might be improvements in sample preparation and analysis methods.

A couple of the key improvements are the derivatization using propionic anhydride to neutralize the highly basic charge of the proteins and block lysine residues and the improvements in MS sensitivity and accuracy which have helped in the analysis of the proteins and their post translational modifications.

Histone analysis overview

- What are histone proteins and why are they important?
- Why are histones so difficult to analyze?
- Iterative search algorithm vs error tolerant searches
- What Mascot server settings that effect modification identification.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



- What are histone proteins and why are they important?
- Why are histones so difficult to analyze?
- Iterative search algorithm vs error tolerant searches
- What Mascot server settings that effect modification identification.

What are histone proteins and why are they important?

- Histones are proteins that contain a lot of basic amino acids (Lys & Arg).
- Main protein component of chromatin in the nuclei of eukaryotic cells.
- DNA is wound around Histone Octamer complexes.
- There are a lot of modifications that have an epigenetic effect on gene expression.
- Involved in many diseases from Alzheimer's and Huntington's to cancer.

MASCOT

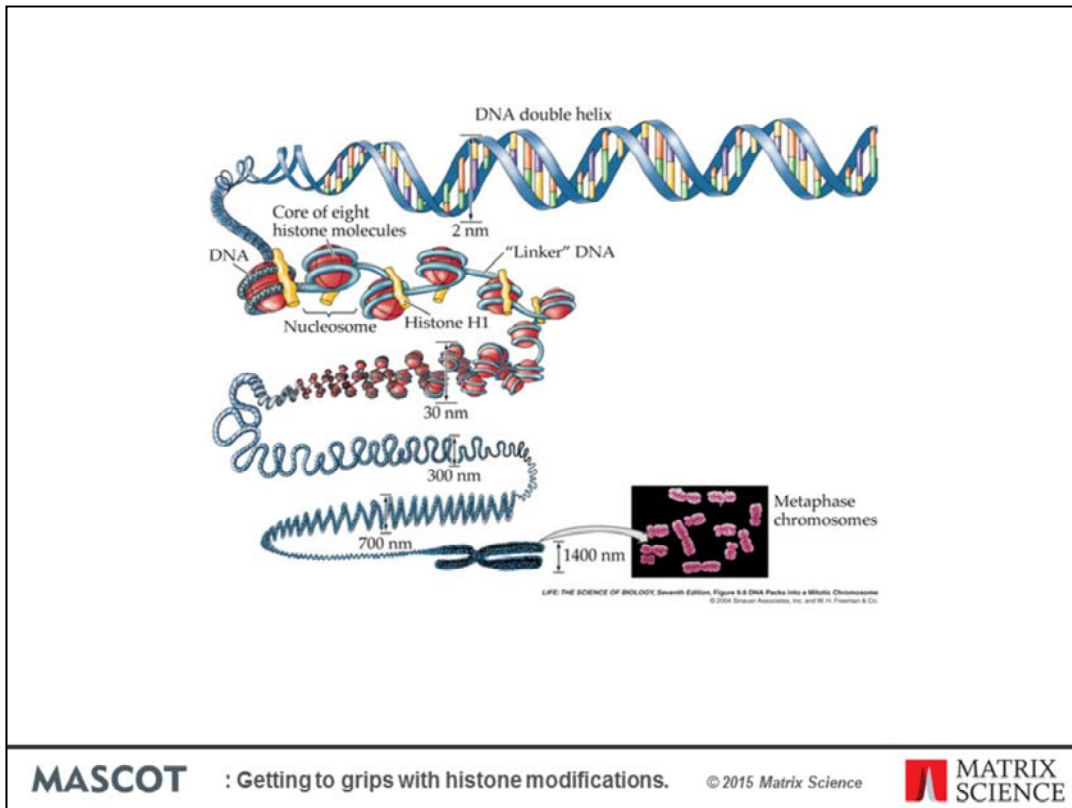
: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

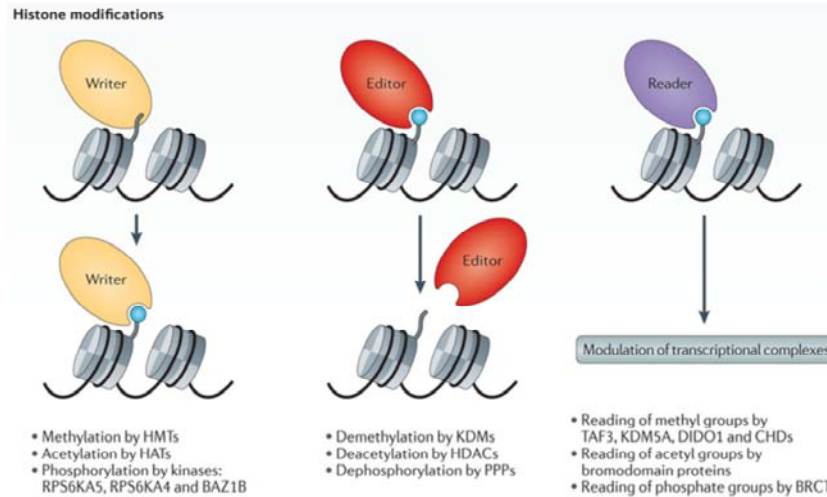
The first question is what are histone proteins and why are they so important?

- Histones are proteins that contain a lot of basic amino acids, lysine and arginine.
- They are the main protein component of chromatin in nuclei of eukaryotic cells.
- DNA is wound around the histone octamer complexes.
- Parts of the histone proteins are exposed to other proteins in the nucleus and there are a lot of modifications that can have an epigenetic effect on gene expression.
- Histones have been shown to be involved in many diseases from Alzheimer's and Huntington's to cancer.



This image shows how DNA is wound around the histone octamers and the DNA histone complexes form a compacting structure that both protects the DNA and allows access for gene expression and replication. Post translational modifications of histones, particularly methylation and acetylation, effect the local chromatin structure.

Histone modifications



Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer
Christoph Plass et al, *Nature Reviews Genetics* 14,765-780 (2013) doi:10.1038/nrg3554

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

**MATRIX
SCIENCE**

The histone modifications are part of what is being called the histone code. Modifications are created or written by a number of transferases and kinases enzymes to the exposed N-terminal tails of the histones. Another set of enzymes can remove or edit these modifications, for example phosphoprotein phosphatases (PPP). The final set of proteins read these modifications and modulate transcription. As these effects on transcription cannot be determined from DNA sequencing we have to use mass spectrometry to determine the modification states of the histones. The code is thought to be more complicated than simple single modifications switches and instead involves multiple modifications and their stoichiometry.

Also see: The complex language of chromatin regulation during transcription. Shelley L. Berger, *Nature* 447, 407-412 (24 May 2007).

<http://www.nature.com/nature/journal/v447/n7143/full/nature05915.html>

So just how many modifications are involved?

- **At least 15 different known modifications**
 - Methylation
 - Acetylation
 - Propionylation
 - Butyrylation
 - Crotonylation
 - Formylation
 - Ubiquitination
 - Citrullination
 - Phosphorylation
 - Hydroxylation
 - O-GlcNacetylation
 - ADP ribosylation
- **On between 25 and 45 different sites depending on the protein isoform**

MASCOT

: Getting to grips with histone modifications.

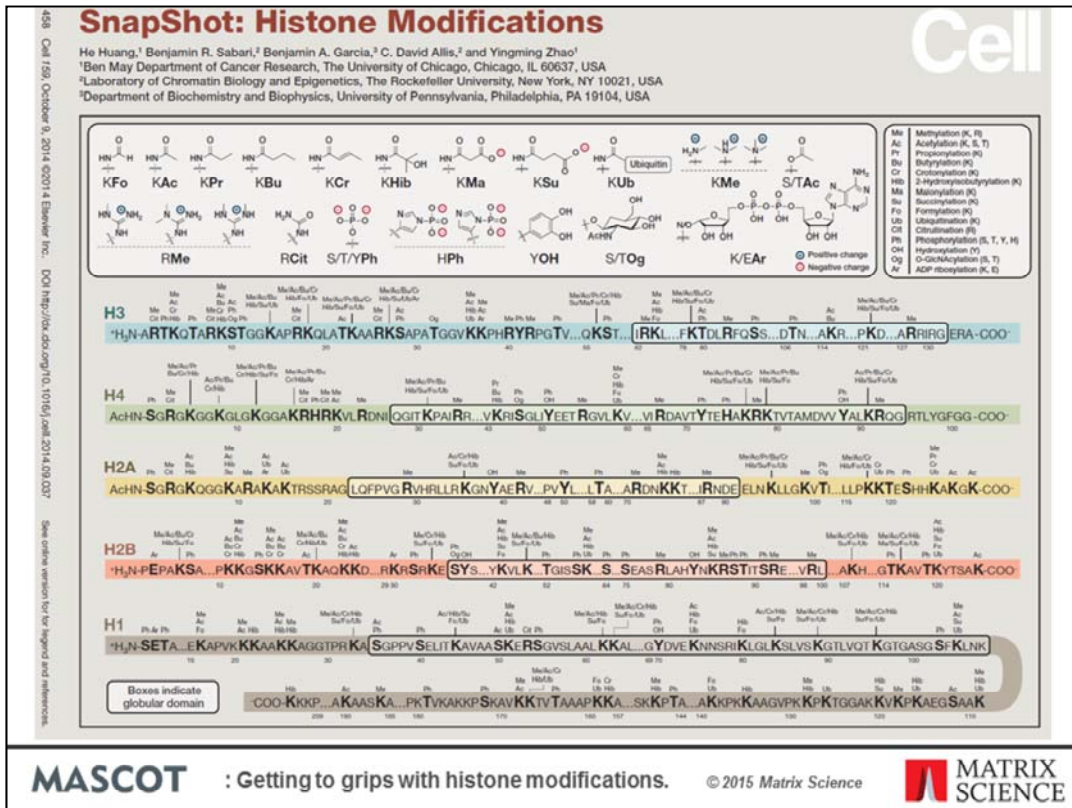
© 2015 Matrix Science

 **MATRIX
SCIENCE**

So just how many modifications are there?

There are at least 15 different known modifications associated with the histone code. The more common ones are methylation, acetylation and propionylation. Less common but important are modifications like phosphorylation.

Current analysis shows that there are between 25 and 45 potential modification sites per histone protein isoform that have been shown to be modified



This snapshot published in Cell shows the extent of the post translational modifications and gives you an idea of the challenges faced during the analysis.

A note about Modification Isoforms

- As far as Mascot Server is concerned Methyl Propionyl, Butyryl and Crotonaldehyde are chemically indistinguishable from one another by chemical composition.
- H(6) C(4) O
- Methyl Propionyl is the combination of a PTM Methylation and the chemical derivatisation of the sample with propionic anhydride.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



A quick note about modification isoforms.

As far as Mascot Server is concerned Methyl Propionyl, Butyryl and Crotonaldehyde are chemically indistinguishable from one another by chemical composition.

The chemical formula is 6 hydrogens 4 carbons and an oxygen.

Methyl Propionyl is the combination of a PTM Methylation and the chemical derivatization of the sample with propionic anhydride. Methyl Propionyl is not included in the list of PSI modifications or the Unimod website, so you can either use one of the two existing modifications or add a new custom modification on your local server.

The effect of high mass accuracy

- The only similar modification that can be distinguished with accurate measurements is Acetyl from Trimethyl.
- However, you can use heavy isotope labeled synthetic peptide standards and retention time information to help assign the modifications on lower resolution data.

J Proteome Res. 2014 Dec 5;13(12):6152-9. doi: 10.1021/pr500902f. Epub 2014 Oct 30.
High resolution is not a strict requirement for characterization and quantification of histone post-translational modifications.
Karch KR, Zee BM, Garcia BA.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

With all these modifications involved you would expect that high mass accuracy will be very important. Many of the modifications or combination of modifications are isobaric so can't be distinguished by accurate mass measurements. The only similar modifications that can be distinguished by accurate measurements is Acetyl from Trimethyl. The mass difference is 0.036385 Da. So, for 2kDa peptide, that would be 18ppm well within the range of a Thermo Orbitrap.

However if you have a lower resolution instrument all is not lost, the lab of Benjamin Garcia showed that you can spike in heavy isotope labeled synthetic peptide standards and use the retention time information to help assign modifications in lower resolution data.

This recent publication caught our eye:

Journal of
proteome
research

Technical Note
pubs.acs.org/jpr

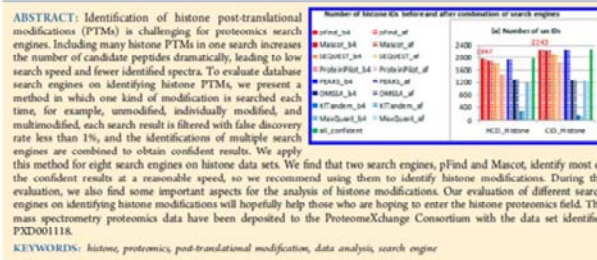
Evaluation of Proteomic Search Engines for the Analysis of Histone Modifications

Zuo-Fei Yuan,[†] Shu Lin,[†] Rosalynn C. Molden,[‡] and Benjamin A. Garcia^{†*}

[†]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, 3400 Civic Center, Building 421, Philadelphia, Pennsylvania 19104, United States

[‡]Department of Chemistry, Princeton University, Washington Road, Princeton, New Jersey 08544, United States

Supporting Information



MASCOT : Getting to grips with histone modifications.

© 2015 Matrix Science



Last year last year this publication caught our eye. The evaluation of Proteomics search engines for the analysis of histone modifications also from the lab of Ben Garcia. The publication used two data sets for the analysis. The first data set was a high resolution and highly accurate HCD data set and the second data set was a more traditional CID data set both acquired on a Thermo Orbitrap.

The paper covers the comparison of different proteomics search engines for these data sets and I'll let you read the paper itself for their conclusions. Instead I decided to use the high resolution HCD data set to compare different PTM analysis strategies using Mascot.

Evaluation of Proteomic search engines data set

- Both high resolution HCD data and lower resolution CID data.
- Propionic anhydride treatment of the sample.
 - No trypsin cleavage at Lys due to the Propionylation - use ArgC as the enzyme
 - Results in longer peptides

ProteomeXchange PXD001118

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

**MATRIX
SCIENCE**

Both the data sets along with search results and analysis are available from ProteomeXchange.

One important thing to note about this data set is that the sample was treated with propionic anhydride. This quantitatively modifies all of the lysine's in the sample such the trypsin cannot cleave at lysine. You could create a custom trypsin enzyme or just use the preexisting ArgC enzyme definition which is what I did.

The propionic anhydride treatment results in longer peptides which is desirable in this case as there are so many lysine and arginine's in the Histones that you would normally end up with very short peptides.

<http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD001118>

Original evaluation of the search engines

- Searched histone only database
- Used iterative search algorithm with multiple searches combining the results

- Search parameters

• Enzyme	:	Arg-C
• Peptide mass tolerance	:	± 10 ppm
• Fragment mass tolerance	:	± 0.02 Da
• Max missed cleavages	:	2
• Instrument type	:	ESI-TRAP
• Fixed modification	:	Propionyl (N-term)

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



In the original evaluation of the search engines the data was searched against a histone only database. The original samples contain many more proteins than just the histones so when we do our analysis we will search against the human taxonomy subset of the SwissProt database and the common contaminants database.

The authors used an iterative search algorithm with multiple searches and combined the results.

The base search parameters used the ArgC enzyme definition, a peptide mass tolerance ± 10 ppm, fragment mass tolerance ± 0.02 Da, maximum number of missed cleavages set to 2 and instrument type ESI trap.

There is also a fixed N-terminal Propionyl modification to account for the sample preparation.

Iterative search algorithm proposed by Huang et al

- Filter unmatched spectra into the next search as per a follow up task.
- Used it to look for adriamycin-induced DNA damage and normal alkylation with side reactions and the iPRG 2011 study data.
- Provided python scripts but only work with OMSSA.

ISPTM: an iterative search algorithm for systematic identification of post-translational modifications from complex proteome mixtures.
Huang Xi, et al, J Proteome Res. 2013 Sep 6;12(9):3831-42. doi: 10.1021/pr4003883.

MASCOT

 **MATRIX
SCIENCE**

The iterative search algorithm that was used is based on a paper by Huang et al. The basic principle is to filter the unmatched spectra from one search into the next search as per a follow-up task in mascot daemon.

Huang and colleagues use it to look for post-translational modifications in a Adriamycin induced DNA damage data set and the data set from the iPRG 2011 study.

The authors used a number of Python scripts to compile the results but the scripts only work with the OMSSA search engine

Iterative search strategy to analyze common Histone modifications

Search title	Modifications
Un	Propionyl K(+56.026)
Ac	Propionyl K(+56.026) Acetyl K (+42.011)
Me	Propionyl K(+56.026) Methyl Propionyl K (70.042)
Di	Propionyl K(+56.026) Dimethyl K (28.031)
Tr	Propionyl K(+56.026) Trimethyl K(42.047)
Ph	Propionyl K(+56.026) Phosphorylation ST (+79.966)
Co	Propionyl K(+56.026), Acetyl K (+42.011), Methyl Propionyl K (70.042), Dimethyl K (28.031), Trimethyl K(42.047), Phosphorylation ST (+79.966)

All search parameters include fixed N-terminal Propionyl.

MASCOT

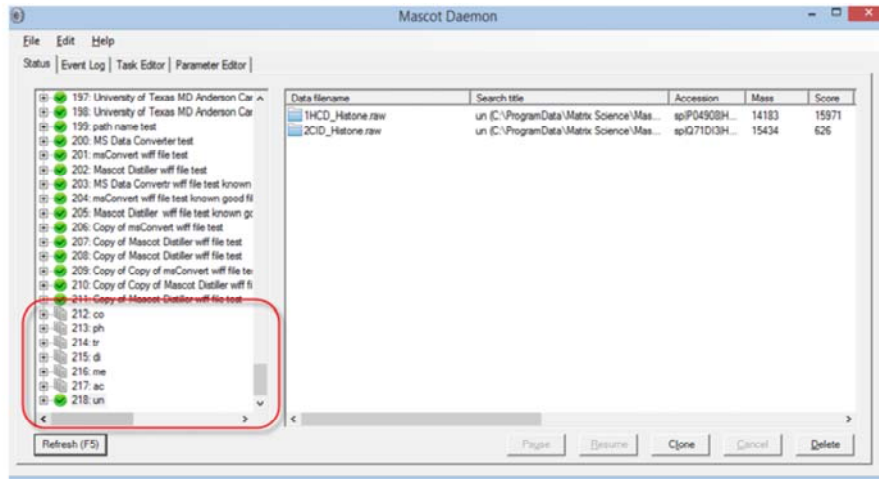
: Getting to grips with histone modifications.

© 2015 Matrix Science

**MATRIX
SCIENCE**

In the comparison analysis this is how the iterative search strategy was set up. The titles stand for Unmodified, Acetyl, Methyl, Dimethyl, Trimethyl, Phosphorylation and Combined searches. You can see the small mass difference between the Acetyl lysine and the Trimethyl lysine which we should be able to differentiate in this experiment.

Can be configured in Mascot Daemon



In this case follow up tasks were configured to search all the spectra at each step as per the paper.

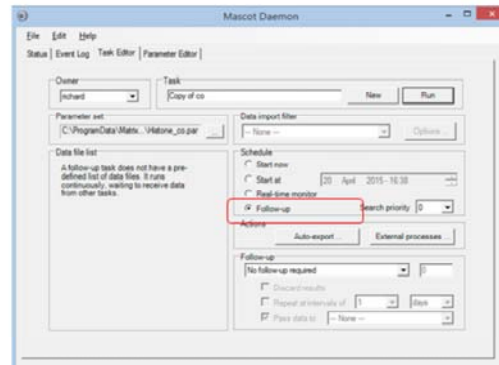
MASCOT

**MATRIX
SCIENCE**

This kind of iterative search strategy can easily be configured in Mascot daemon using follow-up tasks.

How to build a set of iterative searches

- Starting with the last search and build backwards to the first search.



MASCOT : Getting to grips with histone modifications.

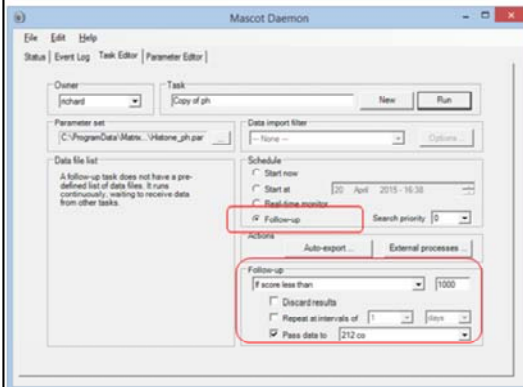
© 2015 Matrix Science



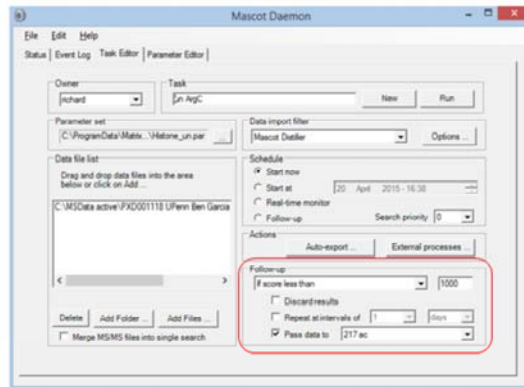
The trick is to start by setting up the last search in the series and build backwards to the first search. In the schedule section of the task editor tab choose follow-up rather than the normal “Start now”.

Middle and “End” tasks

Middle task



Starting task



MASCOT : Getting to grips with histone modifications.

© 2015 Matrix Science



For the middle tasks the schedule has to be both follow-up so that it can receive data from the previous search and you also need to configure the follow-up section so that the task will pass data and to the next search. In the paper all the queries will be passed from one search to the next.

The final task to define is the starting task and this is scheduled to start now but again the follow-up session is configured to pass all the queries through to the next search.

Two approaches to the follow up tasks

Sieve strategy.



Pass unmatched spectra through to the next task, $E > 0.05$.

Multi-search strategy.



Pass through all the queries to each search

MASCOT : Getting to grips with histone modifications.

© 2015 Matrix Science



There are two possible approaches to the follow-up tasks. The standard Mascot method would be to pass the unmatched spectra to the next task, that is queries with a non significant match and expect score of 0.05 or greater. This is like a set of sieves where queries are identified at each search and the unmatched ones go through to the next search.

As we saw in the last slide in the comparison paper they passed all the queries through to the next search. This is effectively the same as automating a set of independent searches and then comparing the results to create a consensus identification for the queries.

Differences between the two methods

Sieve strategy

- **Advantage:**
 - Each query only has one identification so easy to merge results.
 - Can search more modifications than a single search.
- **Disadvantage:**
 - Query may have obtained better results in a later search with different modifications.

Multi-search strategy

- **Advantage:**
 - Each query can find its best match under the search conditions.
- **Disadvantages:**
 - More complicated to combine the results of multiple searches.
 - Final combined search makes earlier searches redundant.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

Both iterative search strategies need to combine their results for the final analysis. In the sieve approach the advantages are that each query only has one significant match so it is easy to merge results. It is not quite true because multiple peptide sequence matches for the query can be used in one report.

Using lot of searches with a small number of modifications in each one allows us to search more modifications in total and each search is more sensitive than a single search with lots of modifications.

The disadvantage is that you don't know if the query would have obtained better results in a later search with different modifications.

For the multiple search approach where all the queries pass-through to the next follow-up task the main advantage is that each query can find its best match under different search conditions. The disadvantages that is more complicated to combine the results of multiple searches. The final combined search makes all the previous steps redundant.

If you wish to compare the effectiveness of a search engines ability to identify a certain PTM this is a good approach. For more general purposes the sieve approach is better.

The traditional Mascot approach is the error tolerant search

- **An error tolerant search was designed to address the following problems:**
 - Enzyme non-specificity
 - Unsuspected chemical & post-translational modifications
 - Peptide sequence not in the database

Error tolerant searching of uninterpreted tandem mass spectrometry data
David M. Creasy and John S. Cottrell, Volume 2, Issue 10, pages 1426-1434,
October 2002

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



At this point I should also mention the traditional Mascot approach to solve this kind of problem which is the error tolerant search. An error tolerant search is a two-part search that was designed to address enzyme non-specificity, unsuspected chemical and post translational modifications and peptide sequences that are not in the database, SNP's for example. Error tolerant searches were incorporated into the mascot search engine over 10 years ago in 2002.

Search constraints for standard, first pass

- Enzyme must be fully specific
- A reduced ceiling on the number of variable modifications
 - default is 2
- Cannot be combined with an automatic decoy database search
- Cannot be combined with quantitation

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



There are a number of constraints for an error tolerant search which prevent the search space from ballooning out of control. For the first pass the enzyme must be fully specific.

We also need to limit the number of variable modifications to two.

we cannot combine an error tolerant search with an automatic decoy database search. Likewise we cannot combine an error tolerant search with quantitation analysis.

Error tolerant, second pass

- The selected enzyme becomes semi-specific
- The complete list of modifications is tested, serially
- All possible amino acid substitutions are tested.
- Only one of the above is allowed per peptide.
- If the mass delta of the modification is less than the smaller of the precursor mass tolerance and the fragment mass tolerance, the modification is rejected.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



In the second pass of the search Mascot makes some changes to the search parameters automatically.

- The selected enzyme becomes semi-specific.
- The complete list of modifications is tested, serially
- All possible amino acid substitutions are tested.
- Only one of the above is allowed per peptide.
- If the mass delta of the modification is less than the smaller of the precursor mass tolerance and the fragment mass tolerance, the modification is rejected.

In the report the error tolerant matches have no expect score

Data format Mascot generic ▾

Precursor _____ m/z

Instrument Default ▾

Error tolerant

#4070 ▶2	482.7460	963.4774	963.4774	0.033 0	35		▶2	■ ■ ■ ■	R.KONYAER.V + Propionyl (K) : [+15.0109 at E6]
#4086 ▶3	483.2379	964.4613	964.4614	-0.12 0	43	7.9e-005	▶1 U ■	R.KONYSER.V + Propionyl (K)	
#4087	965.4686	964.4613	964.4614	-0.058 0	27	0.0031	▶1 U ■	R.KONYSER.V + Propionyl (K)	
#4127 ▶2	484.7853	967.5561	967.5563	-0.19 0	34		▶1	R.GKQGGKAR.A + 2 Propionyl (K) : [-0.9840 at C-term]	
#4150 ▶68	485.2770	968.5395	968.5403	-0.87 0	52	1.6e-005	▶1	R.GKQGGKAR.A + 2 Propionyl (K)	
#4190 ▶42	485.2776	968.5406	968.5403	0.26 0	48		▶1 U ■	R.GKTOGGKAR.A + 2 Propionyl (K) : [+27.0109 at T3]	
#4204 ▶2	323.8544	968.5413	968.5403	1.06 0	44		▶1 U ■	R.GKTOGGKAR.A + 2 Propionyl (K) : [+27.0109 at T3]	
#4204 ▶2	323.8544	968.5413	968.5403	1.06 0	42	0.00017	▶2	R.GKQGGKAR.A + 2 Propionyl (K)	

: Getting to grips with histone modifications.

© 2015 Matrix Science

To run an error tolerant search just click the error tolerant checkbox in the search form.

When you review the results you will easily be able to spot the error tolerant matches because they are missing an expect score.

Use an error tolerant search to see what the common modifications are.

- The Modification statistics section of the protein family report list modifications identified in a search and their frequency.
- Introduced in Mascot Server 2.5.
- Note when analysing a sample which is known to be heavily modified the reported delta mass may be the result of multiple modifications.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

To begin my analysis of the data set I decide to perform an error tolerant search with minimal variable modifications to see what the common modifications were.

- The modification statistic section of the protein family report lists the modifications identified in a search and their frequency.
- It was introduced to Mascot Server version 2.5.
- One thing to be aware of is that when analyzing a sample which is known to be heavily modified the reported mass may be a result of multiple modifications. Even if you haven't used an exotic derivatization or labeling reaction on your sample the combined PTM's may equal the mass of an unusual modification.

Search title : un ArgC SP Histone analysis
Databases : 1: SwissProt 2015_05 (548,454 sequences; 195,409,447 residues)
 2: contaminants 20090624 (262 sequences; 133,770 residues)
Taxonomy : 1: Homo sapiens (human) (20,199 sequences)
 2: (none)
Timestamp : 26 May 2015 at 21:32:36 GMT

All
 Non-significant
 Unassigned
 [\[help\]](#)

 As


Not what you expected? Try [the select summary](#).

▼ Search parameters

Type of search	: MS/MS Ion Search
Enzyme	: Arg-C
Fixed modifications	: Propionyl (N-term)
Variable modifications	: Propionyl (K)
Mass values	: Monoisotopic
Protein mass	: Unrestricted
Peptide mass tolerance	: ± 10 ppm
Fragment mass tolerance	: ± 0.02 Da
Max missed cleavages	: 2
Instrument type	: ESI-TRAP
Number of queries	: 16,906

▶ Score distribution
 ▶ Modification statistics
 ▶ Legend

Protein Family Summary

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science 

As an error tolerant search can not currently be combined with a decoy search. To estimate a suitable significance threshold for the error tolerant search we can run a standard search.

▶ Search parameters
 ▶ Score distribution
 ▶ Modification statistics
 ▶ Legend

Protein Family Summary

Filter Significance threshold p< 0.0020! Max. number of families AUTO [\[help\]](#)
 Ions score or expect cut-off 0 Dendrograms cut at 0
 Show Percolator scores
 Preferred taxonomy All entries

▼ Decoy search summary (reversed protein sequences)

Peptide matches	in target	in Decoy	FDR
- above identity threshold	2024	20	0.99%
- above identity or homology threshold	2062	20	0.97%

Adjust to 1% *

Decoy results are available in [the decoy report](#).

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science **MATRIX SCIENCE**

Adjust the False Discovery Rate to 1% and use the resulting significance threshold with the error tolerant search.

Search title : un ArgC SP Histone analysis Error Tolerant
MS data file : 1HCD_Histone.raw.-1_no_sum.mgf
Databases : 1: SwissProt 2015_05 (548,454 sequences; 195,409,447 residues)
 2: contaminants 20090624 (262 sequences; 133,770 residues)
Error tolerant search of all significant protein families
Taxonomy : 1: Homo sapiens (human) (20,199 sequences)
 2: (none)
Timestamp : 21 May 2015 at 03:35:46 GMT


All
 Non-significant
 Unassigned
 [\[help\]](#)

 As

Not what you expected? Try [the select summary](#).

▼ Search parameters
Type of search : MS/MS Ion Search
Error tolerance : Error tolerant search of all significant protein families
Enzyme : Arg-C
Fixed modifications : [Propionyl \(N-term\)](#)
Variable modifications : [Propionyl \(K\)](#)
Mass values : Monoisotopic
Protein mass : Unrestricted
Peptide mass tolerance : ± 10 ppm
Fragment mass tolerance : ± 0.02 Da
Max missed cleavages : 2
Instrument type : ESI-TRAP
Number of queries : 16,906


► Score distribution
► Modification statistics
► Legend
Protein Family Summary

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science 

Here are the results of a simple search using just fixed Propionyl at the peptide N-terminal and variable propionyl on the lysine. The significance threshold has been set to the value from the standard search.

► Search parameters
 ► Score distribution
 ▼ Modification statistics

Modification	Site	Above thr.	ET	Total matches
Propionyl	K	9892	0	9892
Propionyl	N-term	4040	0	4040
Glu->Gln	E	0	158	158
Propionyl	S	0	154	154
Oxidation	M	0	124	124
Non-specific cleavage	-	0	122	122
Acetyl	K	0	121	121
Delta:H(8)C(6)O(2)	K	0	105	105
Propionyl	T	0	103	103
Guanidinyl	K	0	92	92
Crotonaldehyde	K	0	82	82
Guanidinyl	N-term	0	57	57
Amidated	C-term	0	56	56
Ethyl	N-term	0	56	56
Delta:H(4)C(3)O(1)	C	0	56	56
Ethyl	K	0	54	54
Trimethyl	K	0	42	42
Dehydrated	D	0	39	39
Lys->Xle	K	0	34	34
Asp->Asn	D	0	32	32
Met-loss+Acetyl	N-term	0	30	30
Thr->Gln	T	0	29	29
Gly	K	0	23	23
Dimethyl	R	0	22	22

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science 

Expanding the modification statistics displays an ordered list of the most frequent modifications. You can see the two modifications that were included in the first pass of the search listed at the top of the table. In total there were over 300 different modifications identified.

Not all of the modifications listed in this table effect histones but many of the more frequent ones are important.

Likewise some modifications like oxidized methionine are quite frequent but are not necessarily interesting as they either do not effect Histones or have no known meaning in the Histone code.

We often see a Delta or label reported in the error tolerant results but this does not mean that you should ignore it. The Delta:H(8)C(6)O(2) modification has a mass of 112 which is equivalent to two Propionyl modifications. Likewise Delta:H(4)C(3)O(1) has a mass of 56 so is also equivalent to a single Propionyl modifications of Cysteine which could be a chemical artifact or a result of misassignment due to lack site localization information in the spectra.

Review the modifications

Proteins (128) Report Builder Unassigned (13289)

Protein hits (13 proteins)

Columns: Standard (12 out of 16)

Filters: "Methyl_Propionyl (K)" is in a significant peptide


Methyl_Propionyl (K) is not in a significant peptide Remove

AND Family Update

Export as CSV

#	Family	M	DB	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1		2	SwissProt	f2:H2A11_HUMAN	1712	14712	435	38	16	4	10.57	Histone H2A type 1-J OS=Homo sapiens GN=HST1H2A1 PE=1 SV=3
1		3	SwissProt	f2:H2A1H_HUMAN	1701	14682	416	39	15	4	16.74	Histone H2A type 1-H OS=Homo sapiens GN=HST1H2AH PE=1 SV=3
1		4	SwissProt	f2:H2A2A_HUMAN	1221	14927	398	31	16	4	56.57	Histone H2A type 2-A OS=Homo sapiens GN=HST2H2AA3 PE=1 SV=3
2		1	SwissProt	f2:H4_HUMAN	2193	12033	618	137	19	9	2581.42	Histone H4 OS=Homo sapiens GN=HST1H4A PE=1 SV=2
2		1	SwissProt	f2:H14_HUMAN	1635	25382	96	41	4	3	3.41	Histone H1.4 OS=Homo sapiens GN=HST1H1E PE=1 SV=2
2		2	SwissProt	f2:H12_HUMAN	1617	24713	87	43	3	3	4.91	Histone H1.2 OS=Homo sapiens GN=HST1H1C PE=1 SV=2
4		1	SwissProt	f2:H32_HUMAN	1497	16163	528	93	19	7	402.98	Histone H3.2 OS=Homo sapiens GN=HST2H3A PE=1 SV=3
4		2	SwissProt	f2:H33_HUMAN	644	16103	458	54	15	7	134.18	Histone H3.3 OS=Homo sapiens GN=H3F3A PE=1 SV=2
2		1	SwissProt	f2:RL29_HUMAN	343	19590	79	20	10	5	11.54	60S ribosomal protein L29 OS=Homo sapiens GN=RPL29 PE=1 SV=2
20		1	SwissProt	f2:H2B1H_HUMAN	154	15060	129	8	12	1	1.23	Histone H2B type 1-H OS=Homo sapiens GN=HST1H2BH PE=1 SV=3
42		1	SwissProt	f2:RL36A_HUMAN	129	13721	13	3	6	2	1.38	60S ribosomal protein L36a OS=Homo sapiens GN=RPL36A PE=1 SV=2
53		1	SwissProt	f2:H2B1L_HUMAN	104	15120	81	5	10	1	1.23	Histone H2B type 1-L OS=Homo sapiens GN=HST1H2BL PE=1 SV=3
95		1	SwissProt	f2:H2BFS_HUMAN	77	15056	57	1	9	1	0.49	Histone H2B type F-S OS=Homo sapiens GN=H2BFS PE=1 SV=2

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science



At this point it is worth while reviewing the search results and modification matches..

The quick way to do this is by using the Report builder tab of the protein family report. Here I filter on the Methyl_Propionyl modification and see that apart from the 60s ribosomal proteins it is only found on Histone proteins.

Proteins (201) Report Builder Unassigned (12632) [Help](#)

Protein hits (45 proteins)

Columns: Standard (12 out of 16)


Filters: "Oxidation (M)" is in a significant peptide

Oxidation (M) is in a significant peptide Remove

AND Family Update

Export as CSV

Family	#	ID	Accession	Score	Mass	Matches	Match(sig)	Sequences	Seq(sig)	emPAI	Description
1	1	SwissProt	f2:1H_HUMAN	21576	12033	618	582	19	18	553756451.13	Histone H4 OS=Homo sapiens GN=HIST1H4A PE=1 SV=2
2	8	SwissProt	f2:1H2AC_HUMAN	7750	14820	281	242	15	12	4970.36	Histone H2A type 2-C OS=Homo sapiens GN=HIST2H2AC PE=1 SV=4
3	1	SwissProt	f2:1H3_HUMAN	13630	16163	528	425	19	16	65927798.67	Histone H3.2 OS=Homo sapiens GN=HIST3H2A PE=1 SV=3
3	2	SwissProt	f2:1H31_HUMAN	13461	16179	521	421	19	16	31136604.74	Histone H3.1 OS=Homo sapiens GN=HIST3H3A PE=1 SV=2
3	3	SwissProt	f2:1H31T_HUMAN	10953	16283	395	312	15	11	319376.22	Histone H3.1T OS=Homo sapiens GN=HIST3H3 PE=1 SV=3
3	4	SwissProt	f2:1H33_HUMAN	10429	16103	458	365	15	12	796101.96	Histone H3.3 OS=Homo sapiens GN=H3F3A PE=1 SV=2
2	1	SwissProt	f2:1L29_HUMAN	2525	19590	79	65	10	9	295.10	60S ribosomal protein L29 OS=Homo sapiens GN=RPL29 PE=1 SV=2
8	2	SwissProt	f2:1H2BK_HUMAN	2225	15058	107	90	12	10	1994.78	Histone H2B type 1-K OS=Homo sapiens GN=HIST2H2BK PE=1 SV=3
9	1	SwissProt	f2:1CCD6_HUMAN	1784	42228	70	55	22	20	38.18	Coiled-coil domain-containing protein 86 OS=Homo sapiens GN=CCDC86 PE=1 SV=1
10	1	SwissProt	f2:1L26_HUMAN	1703	18648	55	51	12	11	527.23	60S ribosomal protein L26 OS=Homo sapiens GN=RPL26 PE=1 SV=1
10	2	SwissProt	f2:1L26L_HUMAN	1417	18646	42	38	9	8	100.46	60S ribosomal protein L26-like 1 OS=Homo sapiens GN=RPL26L1 PE=1 SV=1
13	1	SwissProt	f2:1L36_HUMAN	1244	13254	39	32	11	10	2124.82	60S ribosomal protein L36 OS=Homo sapiens GN=RPL36 PE=1 SV=3
16	1	SwissProt	f2:1HNPCL_HUMAN	1041	35330	36	27	8	7	6.35	Heterogeneous nuclear ribonucleoproteins C1/C2 OS=Homo sapiens GN=HNPCL PE=1 SV=4
17	1	SwissProt	f2:1L13_HUMAN	1006	25703	33	28	13	13	29.64	60S ribosomal protein L13 OS=Homo sapiens GN=RPL13 PE=1 SV=4
18	1	SwissProt	f2:1OHTOP_HUMAN	870	27109	23	20	7	7	9.29	Chromatin target of FRMT1 protein OS=Homo sapiens GN=OHTOP PE=1 SV=2
22	1	SwissProt	f2:1S19_HUMAN	716	16947	22	20	7	7	106.00	40S ribosomal protein S19 OS=Homo sapiens GN=RPS19 PE=1 SV=2
23	1	SwissProt	f2:1L13A_HUMAN	698	25187	24	23	8	7	11.12	60S ribosomal protein L13a OS=Homo sapiens GN=RPL13A PE=1 SV=2
29	1	SwissProt	f2:1EBP2_HUMAN	598	37352	28	22	12	11	6.30	Probable rRNA-processing protein EBP2 OS=Homo sapiens GN=EBP2 PE=1 SV=2
20	1	SwissProt	f2:1THOCA_HUMAN	589	27544	20	17	8	8	8.93	THO complex subunit 4 OS=Homo sapiens GN=ALYREF PE=1 SV=3
21	1	SwissProt	f2:1L19_HUMAN	582	25300	40	30	13	12	31.42	60S ribosomal protein L19 OS=Homo sapiens GN=RPL19 PE=1 SV=1
23	1	SwissProt	f2:1S28_HUMAN	505	8060	15	15	7	7	242.66	40S ribosomal protein S28 OS=Homo sapiens GN=RPS28 PE=1 SV=1
25	1	SwissProt	f2:1HOKO_HUMAN	489	55766	12	10	6	5	1.26	Non-POU domain-containing octamer-binding protein OS=Homo sapiens GN=HOKO PE=1 SV=4
28	1	SwissProt	f2:1L7_HUMAN	456	31168	13	13	7	7	4.11	60S ribosomal protein L7 OS=Homo sapiens GN=RPL7 PE=1 SV=1
53	1	SwissProt	f2:1NUCKS_HUMAN	403	29465	13	11	3	3	1.37	Nuclear ubiquitin casein and cyclin-dependent kinase substrate 1 OS=Homo sapiens GN=NUCKS1 PE=1 SV=1
59	1	SwissProt	f2:1S14_HUMAN	379	18991	12	12	4	4	7.64	40S ribosomal protein S14 OS=Homo sapiens GN=RPS14 PE=1 SV=3
52	1	SwissProt	f2:1L39_HUMAN	344	6963	18	18	3	3	104.21	60S ribosomal protein L39 OS=Homo sapiens GN=RPL39 PE=1 SV=2
55	1	SwissProt	f2:1L21_HUMAN	325	19954	12	8	7	5	3.70	60S ribosomal protein L21 OS=Homo sapiens GN=RPL21 PE=1 SV=2

MASCOT : Getting to grips with histone modifications. © 2015 Matrix Science 

While MetOx is distributed evenly through the protein hits with about one third of the proteins containing an oxidized Met.

If we look at the top Histone hit by clicking on the protein accession number

#10895	80 - 93	881.9936	1761.9726	1761.9699	1.55	0	88	5.4e-009	1	U	R.KTVTAMDVVYALKR.Q
#10896	80 - 93	881.9944	1761.9742	1761.9699	2.45	0	105	1.8e-010	1	U	R.KTVTAMDVVYALKR.Q
#10897	80 - 93	881.9948	1761.9751	1761.9699	2.93	0	69	3.6e-007	1	U	R.KTVTAMDVVYALKR.Q
#10898	80 - 93	881.9967	1761.9787	1761.9699	5.02	0	52	1.4e-005	1	U	R.KTVTAMDVVYALKR.Q
#11032	80 - 93	889.9891	1777.9636	1777.9648	-0.67	0	114	5.9e-011	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11033	80 - 93	593.6619	1777.9638	1777.9648	-0.55	0	76	7.4e-008	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11034	80 - 93	593.6619	1777.9639	1777.9648	-0.54	0	90	4.4e-009	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11035	80 - 93	593.6620	1777.9642	1777.9648	-0.35	0	74	1.2e-007	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11036	80 - 93	593.6621	1777.9644	1777.9648	-0.25	0	84	1.3e-008	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11037	80 - 93	889.9897	1777.9648	1777.9648	0.011	0	95	1.3e-009	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11038	80 - 93	593.6622	1777.9649	1777.9648	0.036	0	77	6.3e-008	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11039	80 - 93	593.6623	1777.9651	1777.9648	0.17	0	71	2.2e-007	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11040	80 - 93	593.6631	1777.9676	1777.9648	1.54	0	70	2.5e-007	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0)
#11262	80 - 93	607.0061	1817.9964	1817.9961	0.17	0	16	0.033	1	U	R.KTVTAMDVVYALKR.Q + Propionyl (T)
#11263	80 - 93	607.0062	1817.9968	1817.9961	0.39	0	13	0.067	1	U	R.KTVTAMDVVYALKR.Q + Propionyl (T)
#11264	80 - 93	910.0065	1817.9985	1817.9961	1.29	0	28	0.024	1	U	R.KTVTAMDVVYALKR.Q + Propionyl (T)
#11265	80 - 93	910.0067	1817.9989	1817.9961	1.52	0	28	0.0037	1	U	R.KTVTAMDVVYALKR.Q + Propionyl (T)
#11266	80 - 93	910.0068	1817.9990	1817.9961	1.59	0	39	0.0017	1	U	R.KTVTAMDVVYALKR.Q + Propionyl (T)
#6292	80 - 93	612.3331	1833.9974	1833.9910	-7.45	0	45	8.7e-005	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)
#11335	80 - 93	612.3373	1833.9901	1833.9910	-0.51	0	18	0.02	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)
#11336	80 - 93	612.3382	1833.9928	1833.9910	0.94	0	13	0.073	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)
#11337	80 - 93	612.3385	1833.9936	1833.9910	1.37	0	49	2.6e-005	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)
#6293	80 - 93	612.3386	1833.9940	1833.9910	1.63	0	15	0.036	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)
#11338	80 - 93	612.3392	1833.9956	1833.9910	2.50	0	15	0.037	1	U	R.KTVTAMDVVYALKR.Q + Oxidation (0); Propionyl (T)

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



We jump to the list of matches. We can see that pretty much all of the peptides are represented in both oxidized and unoxidised forms. This means that although we will increase the number of hits in the search we will probably not add any new sequences, with their biologically relevant modifications, that we did not already know about. The main benefit of including these matches in a search will be to prevent these queries from being misassigned and to help keep the FDR rate numbers reasonable.

Error tolerant search with more variable modifications

- Increase the number of MaxEtVarMods from 2 to 8 in the Configuration Editor
- Choose 5 of the most important Histone modifications:
 - Acetyl(K), Dimethyl(K), Phospho(ST), Propionyl(K) and Trimethyl(K)

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



Can we run an error tolerant search with more variable modifications? Baring in mind mascot can only find one unsuspected modification per a query wouldn't it be better to run a search with more variable modifications in the first pass search? Yes it is possible to increase a configuration value max ET var mods from two modifications in the first pass search up to a maximum of eight. You can edit this setting in the configuration editor. I choose the five modifications used in the combined search as my initial variable modifications. When these settings are used in a standard search completes successfully in a reasonable time frame but when run as an error tolerant search it did not complete.

What happened? A Combinatorial explosion!

- **Histone H4 fragment 5 - 18**
 - GKGGKGLGKGGAKR
 - 5 possible Lys modification states at four Lys modification sites
 - 5^4 permutations = 625 possible arrangements
- **Histone H2A type 2A fragment 73 -130**
 - DNKKTRIIPRHLQLAIRNDEELNKLLGKVTIAQGGVLPNIQAVLLPKKTESHHKAKGK
 - 5 possible states with 8 Lys and 2 Phospho states at 4 Ser or Thr sites
 - $5^8 \times 2^4$ permutations = $390625 \times 16 = 6,250,000$ arrangements
- **The error tolerant search then adds all the unsuspected modifications on top of these calculations.**

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

What happened? The search suffered from a combinatorial explosion.

Here is an example of a small peptide from histone H4. From the variable modifications selected in the first pass search there are five possible lysine modification states at four different lysines in the peptide.

The number of permutations and combinations calculates out to 625 possible arrangements. This has increased the search space a considerable amount but it is still possible to search the expanded space in a timely fashion.

Here is another longer peptide fragment from histone2A type 2A. Again in the search we had five possible variable lysine modification states and there are 8 lysine's in the peptide. There is also a chance of phosphorylation. There are two possible phosphorylation states phosphorylated or non-phosphorylated and four possible sites for either serine and threonine phosphorylation. This calculates out to over 6 million possible arrangements.

Mascot Server then has to add a layer of error tolerant modifications over the top of this so each one of the possible permutations and combinations is modified with each of the modifications defined on the server.

Once you consider that there may be many thousands of long peptides like this in the database you can see how Mascot Server is going to run out of resources when searching with this many variable modifications.

Order is important and repetition (same modification at different sites is allowed) so the permutation formula is n^r .

Lys modifications and phosphorylation modifications are independent events so we can multiple the two permutation factors together.

We cannot perform an error tolerant search of all of the modifications that we expect to observe. However we can probably search with less variable modifications and still obtain satisfactory results.

Alternatively we can try iterative searches.

With a longer peptide and ppm accuracy, more of the queries will match each possible peptide combination.

Sieve strategy vs multi-search strategy vs error tolerant with modification counts.

	Iterative searches - Sieve	Iterative searches - Multi-search	Error tolerant search
No. Queries	16906	16906	16906
un (Propionyl(K))	2680	2680	4040 (9892)
Ac (Acetyl(K))	84	86	121
Me (Methyl_Propionyl(K))	178	242	82
Di (Dimethyl(K))	95	117	0
Tr (Trimethyl(K))	38	69	42
ph(S) (Phospho(S))	12	13	7
ph(T) (Phospho(T))	1	3	0
co	375	1666	-
Total	3463	4876	5764

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

MATRIX SCIENCE

The iterative search strategies were set up in Mascot daemon and a peak list was passed down the chain of tasks.

One set of tasks were set up to analyze the data using the sieve approach.

A second set of tasks passed all the queries through to the next task after each step.

Finally an error tolerant search was run with the using the same settings as the Un step, fixed Propionyl at the N-terminal and variable Propionyl(K).

All searches were adjusted to 1%FDR or as close to 1% as possible.

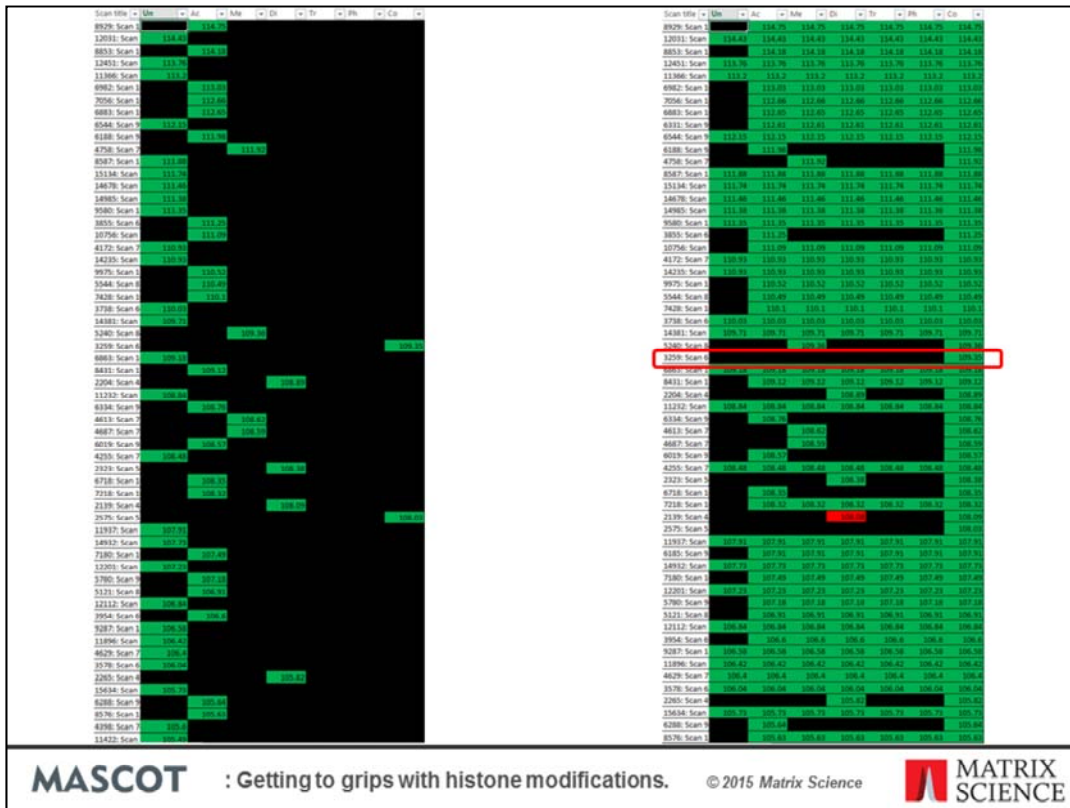
There were 9892 Propionyl(K) modification sites but many peptides have more than one site of modification so the total numbers of identified peptide will be lower than these numbers.

The table shows the number queries identified with the modifications being tested in the search.

I totaled the number of modifications identified in the iterative search steps and compared it to the number of queries identified by the combined search and they were pretty even.

The error tolerant search identifies a lot more queries with at least one Propionyl(K)

and another unsuspected modification. However the numbers quickly drop off for the other interesting modifications because we have more than one unsuspected modification per a peptide.



The one difficulty with analyzing the iterative searches is compiling lists of the matching peptides and comparing the results. After a bit of Excel trickery and use of context highlighting we can compare the matches to the sieve style iterative approach on the left to the multi-search iterative approach. The highest scoring queries are in green. Lower scoring matches are in red and matches with a score less than 20 or were not found in the search are black.

The sieve approach shows that queries were only used once per a search with a sparse array of matches.

The multi-search iterative approach shows how queries that match a peptide with Propionyl at the N-terminal and variable Propionyl(K) are propagated through all the following searches.

Looking through the results of the multi-search iterative strategy we can see examples where a query has only matched a peptide under those specific search conditions or in some cases different peptide under one set of search conditions.

Multi-search strategy advantage

4490b: Scan	53.53	7471: Scan 1	71.89	71.89	71.89	71.89	71.89	71.89	71.89
9025: Scan 1	53.53	8585: Scan 9	71.85	71.85	71.85	71.85	71.85	71.85	71.85
4713: Scan 2	53.53	3678: Scan 6	71.82	71.82	71.82	71.82	71.82	71.82	71.82
8924: Scan 1	53.53	8924: Scan 1	53.53	71.59	53.53	53.53	53.53	53.53	71.59
7335: Scan 1	53.5	9721: Scan 1	71.5	71.5	71.5	71.5	71.5	71.5	71.58
10361: Scan	53.48	10758: Scan	71.49	71.49	71.49	71.49	71.49	71.49	71.49
8350: Scan 1	53.47	1256: Scan 3	71.49	71.49	71.49	71.49	71.49	71.49	71.49
		7158: Scan 1	71.43	71.43	71.43	71.43	71.43	71.43	71.43
		16093: Scan	71.43	71.43	71.43	71.43	71.43	71.43	71.43
		6457: Scan 9	71.42	71.42	71.42	71.42	71.42	71.42	71.42
		8098: Scan 1	71.39	71.39	71.39	71.39	71.39	71.39	71.39

Search	Score	Sequence	Mods
Ac, Co	72	PEPAKSAPAPKKGSKKAVTKAQKKGDKKR	Acetyl (K); 9 Propionyl (K)
The rest	54	PEPAKSAPAPKKGSKKAVTKAQKKGDKKR	10 Propionyl (K)

```

8924: Scan 12566 (rt=34.9408) (C:\MSData active\FXD001118 UPenn Ben Garcia Histone\1MCD_Histone.raw)
Score > 20 indicates identity
#15383      1221.3673 3661.0801 3661.0817  -0.46 0 72 3.5e-007 1 U ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 72 3.5e-007 1 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 61 4.1e-006 3 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 58 7.9e-006 4 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 54 2.3e-005 5 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + 10 Propionyl (K)
          -0.46 0 51 4e-005 6 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 30 0.0049 7 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
          -0.46 0 26 0.012 8 ■ M.PEPAKSAPAPKKGSKKAVTEVQKKGDKKR.K + 2 Acetyl (K); 8 Propionyl (K)
          -0.46 0 26 0.014 9 ■ M.PEPAKSAPAPKKGSKKAVTEVQKKGDKKR.K + 2 Acetyl (K); 8 Propionyl (K)
          -0.46 0 15 0.15 10 ■ M.PEPAKSAPAPKKGSKKAVTKAQKKGDKKR.K + Acetyl (K); 9 Propionyl (K)
  
```

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



Also in the multi-search iterative strategy results we can find queries that have a different score for different searches. In this example a query has a significant match in the first Un search. In the sieve search strategy the query is filtered out from the subsequent searches. In the multi-search strategy the query scores higher in the Acetyl and Combined searches. The query is matching the same sequence in all searches but with a slightly different set of modifications, an acetyl and 9 Propionyl instead of 10 Propionyl. The score difference is significantly higher too. If we click on the query number in the report we can open the peptide view and at the bottom of the page see the top 10 hits for the search and the site analysis results.

We know this problem exists and leads to an increased FDR but it is very infrequent. If you wish to dig deeper into the results click on the query number to open the peptide view.

Site determination

All matches to this query

Score	Mr(calc)	Delta	Sequence	Site Analysis
71.6	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K12, K15, K16, K20, K23, K24, K27, K28, Acetyl K11; 46.73%
71.6	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K11, K15, K16, K20, K23, K24, K27, K28, Acetyl K12; 46.73%
60.9	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K11, K12, K16, K20, K23, K24, K27, K28, Acetyl K15; 4.01%
58.1	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K11, K12, K15, K16, K20, K23, K24, K27, K28, Acetyl K5; 2.10%
53.5	3661.0818	-0.0017	PDPAKSAPAPKKGSKKAVTKAQKKGKRR	
51.1	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K11, K12, K15, K20, K23, K24, K27, K28, Acetyl K16; 0.42%
30.1	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K11, K12, K15, K16, K23, K24, K27, K28, Acetyl K20; 0.00%
26.3	3661.0818	-0.0017	PDPAKSAPAPKKGSKKAVTKVQKKGKRR	
25.6	3661.0818	-0.0017	PDPAKSAPAPKKGSKKAVTKVQKKGKRR	
15.4	3661.0817	-0.0017	PEPAKSAPAPKKGSKKAVTKAQKKGKRR	Propionyl K5, K11, K12, K15, K16, K20, K24, K27, K28, Acetyl K23; 0.00%

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

**MATRIX
SCIENCE**

At the bottom of the peptide view is the top ten list of matches for a query. If there are multiple possibilities for the localization of the modifications Mascot will carry out a site analysis. The Site analysis is not just for phosphor peptides. Mascot Server will report the potential site assignments for the top 10 matches with identical modifications. Here we can see the two most likely sites for the Acetyl PTM. The alternative peptide match with 10 Propionyl is here in the middle of top 10 matches. Although it does not have a site assignment because it has different modifications we can tell by the neighboring assignments which have a probability of less than 1% it is an unlikely match.

Refine the search strategy

- **Too many unsuspected modifications per a peptide for an error tolerant search.**
- **Multi Search strategy as described to evaluate the search engines can be replaced by a single combined search.**
- **Sieve strategy**
 - Select different common modifications to see which ones are the most important.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science



As there are too many unsuspected modifications per a peptide for an error tolerant search. The multi-search strategy that was used to evaluate the search engines can be replaced by a single combined search. Adding more variable modifications to the search parameters increases the identity threshold such you lose more matches than you gain.

This makes the combined search parameters, with the change of Propionyl (K) to a fixed mode the sweet spot as far as number of variable mods goes and biologically relevant information.

We can obtain more matches by expanding the Sieve strategy to include some of the non biologically relevant modifications to increase the total number of matches. We will pick up quite a few biologically relevant peptides too

Refining the sieve strategy

- **Make Propionyl (K) a fixed modification like Propionyl (N-term)**
 - Reduces the search space as we only look for modified lysines.
- **Which modifications to use?**
 - Propionyl (S), Propionyl (T), Oxidation (M)

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

The initial error tolerant search reported a number of matches with unmodified lysine sites but these are not really what we are interested in. Instead let's make the Propionyl (K) a fixed modification. This means that Mascot Server will only search for modified lysines with either Propionyl (K) or one of the variable Lysine modifications.

Next we will expand the number of search iterations we will use by adding some of the most frequently occurring modifications to the strategy.

Original sieve strategy vs improved sieve strategy with modification counts.

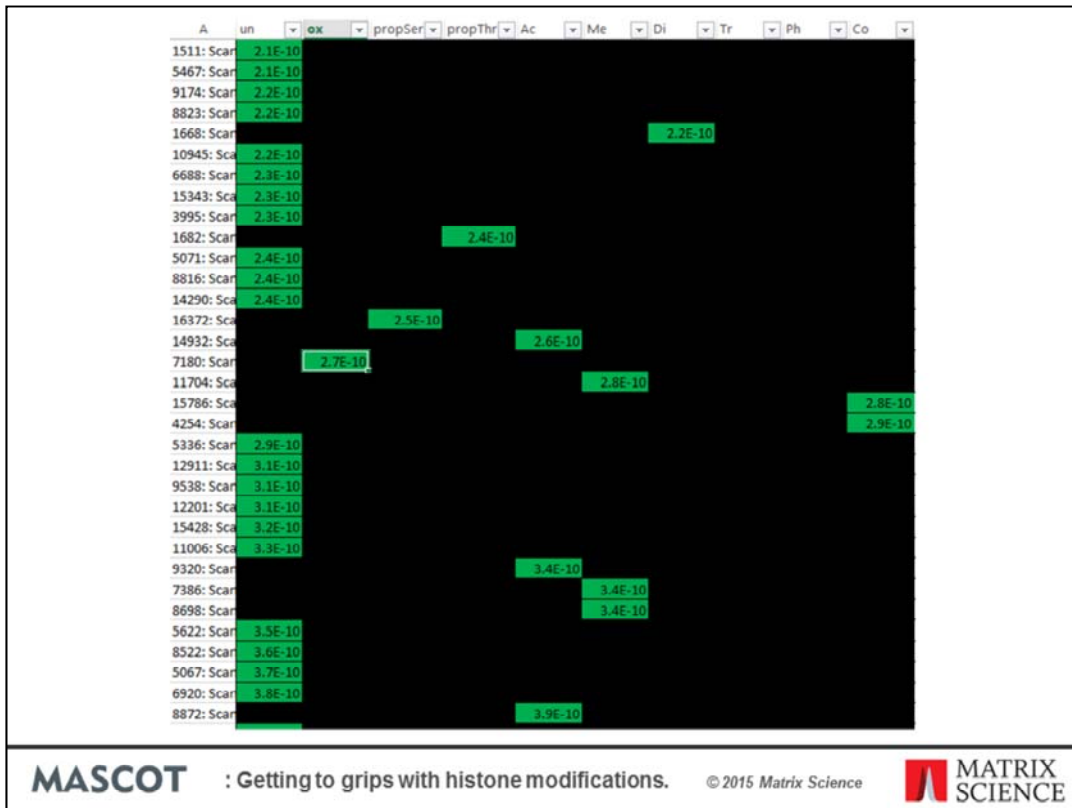
	Iterative searches -Sieve	Iterative searches - improved Sieve
No. Queries	16906	16906
un (Propionyl(K))	2680	4484
Ox (MetOx)	-	130
propSer (Propionyl(S))	-	187
propThr (Propionyl(T))	-	83
Ac (Acetyl(K))	84	90
Me (Methyl_Propionyl(K))	178	79
Di (Dimethyl(K))	95	97
Tr (Trimethyl(K))	38	38
ph(S) (Phospho(S))	12	6
ph(T) (Phospho(T))	1	1
co	375	689
Total	3463	5884

MASCOT : Getting to grips with histone modifications.

© 2015 Matrix Science



After making Propionyl (K) a fixed modification and adding three more rounds of iterative searches we identified more modifications for all but the Methyl_Propionyl search with the refined Sieve strategy. The two big gains in number of modifications were the number of peptides identified in the first and last iterative steps.



I plotted the expect scores for the searches and compared the results. As you can see it will be easy merge the results as there are no conflicting matches.

Search setting recommendations

- Search against a database that contains all the potential proteins in the sample.
- Set both Propionyl (K) and Propionyl (N-term) as fixed modifications.
- Error tolerant search with minimal modifications to determine abundant PTM's
- Use a sieve strategy to identify the most abundant modifications in the sample.

MASCOT

: Getting to grips with histone modifications.

© 2015 Matrix Science

 **MATRIX
SCIENCE**

Search against a database that contains all the potential proteins in the sample. Otherwise you haven't given the data a sufficient statistical challenge.

If you are using chemical derivatization such as propionic anhydride then I recommend setting both Propionyl (K) and Propionyl (N-term). The use of propionic anhydride does have some side reactions and a middle down approach using GluC might be a good alternative.

Run an error tolerant search with minimal modifications first to determine the most abundant modifications.

Finally use a sieve strategy to identify combinations of modified peptides.