Getting to grips with histone modifications.

Richard Jacob
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 analyse?
• Iterative search algorithm vs error tolerant searches
• What Mascot server settings that effect modification identification.

• 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.
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 the 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 of 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.
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.


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

- H(6) C(4) O
- Methyl Propionyl is the combination of a PTM Methylation and the chemical derivatisation of the sample with propionic anhydride.

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.
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 with in 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.
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.
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
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.
The iterative search algorithm that was used is based on a paper by Huang at 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 Adriamycin-induced DNA damage and normal alkylation with side reactions and the iPRG 2011 study data.

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

<table>
<thead>
<tr>
<th>Search title</th>
<th>Modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un</td>
<td>Propionyl K(+56.026)</td>
</tr>
</tbody>
</table>
| Ac           | Propionyl K(+56.036)  
|              | Acetyl K (+42.011)   |
| Me           | Propionyl K(+56.026)  
|              | Methyl Propionyl K (70.042) |
| L1           | 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) |
| Cn           | Propionyl K(+56.026)  
|              | Acetyl K (+52.011)    
|              | 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.
This kind of iterative search strategy can easily be configured in Mascot daemon using follow-up tasks.
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”.

How to build a set of iterative searches

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

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.
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.
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.
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.
- 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.

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

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.
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.
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.
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.
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.
The initial error tolerant search repotted 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.
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 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 proponic anhydride then I recommend setting both Propionyl (K) and Propionyl (N-term). The use of proponic 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.