# Understanding the Pro Group™ Algorithm


RUO-IDV-06-1661-C

AB Sciex Pte. Ltd.
Blk 33, #04-06
Marsiling Ind Estate Road 3
Woodlands Central Indus. Estate
SINGAPORE 739256

## Contents

# Introduction

This document explains the protein grouping issue and how the Pro Group™ algorithm in ProteinPilot™ Software addresses it.

Information in this document addresses the following questions:

- What is the protein grouping issue?
- What kinds of false proteins are reported by software that does not do proper protein grouping?
- How does the Pro Group algorithm prevent the reporting of these false or suspect proteins?
- What is a protein group? What are “competitor” proteins, and how does the ProteinPilot software show them?
- I was looking for a particular protein. Why did I not see it in the results?
- How can I tell when other protein identification software is reporting an invalid number of proteins because of the failure to do proper protein grouping?

This document is intended to help users understand the philosophy behind the Pro Group algorithm. Refer to the ProteinPilot software Help for additional information.

# The Protein Grouping Issue

In what is commonly referred to as “bottom-up” proteomics, intact proteins are not separated before digestion. As a result, the connection between a protein molecule parent and the peptides produced by digestion is lost. The “protein grouping issue” or the “protein inference issue” refers to the need for an analysis after peptide identification to determine which proteins should be reported.

If each MS/MS fragmentation spectrum in the data was associated with only a single protein, this would not be an issue. There are two reasons that this is not that simple:

- **Protein ambiguity**: Any single peptide sequence might be found within multiple protein sequences.
- **Peptide ambiguity**: Search engines might report multiple possible peptides for each fragmentation spectrum.

Because each spectrum can provide evidence for multiple proteins, it is not obvious how to infer which proteins should really be reported. There is, however, an obvious rule that should be followed to make these inferences:

> *You cannot use the same data multiple times to justify the detection of multiple proteins.*

Many current protein identification software tools fail to enforce this rule. The consequence of this failure is that false and redundant proteins get reported as confident identifications, which makes the number of proteins reported falsely high. Most current software does not count obviously redundant proteins as separate detections, but there are many subtle causes of redundancy that are missed. Because such software provides little or no indication of which proteins are most suspect, scientists can unknowingly report significantly inflated protein identification numbers.

The proteomics community has now recognized the importance of this issue. It is one of the major points of the publication guidelines proposed in the journal *Molecular and Cellular Proteomics*[1] (referred to as the "MCP guidelines"). These guidelines are a mandate to see beyond who reported the most proteins and take steps to assure that each reported protein is actually defensible. The guidelines have been crafted in conjunction with, and adopted by, other journals as well, so it is important to understand this issue before publishing results.

It is possible to find and filter out most redundant proteins with other software if the correct software settings are selected and several steps of manual review are performed. It is preferable if the software can do this automatically. The required manual review is difficult to do correctly, and is too time-consuming to keep pace with the rate at which results are generated in high-throughput proteomics.

The key to assuring that each reported protein is actually justified is to determine which groups of proteins derive evidence from largely the same spectra. Each protein group should be analyzed to determine which proteins are proven detected by the data and which are redundant. Determining redundancy is not simple, but the Pro Group algorithm handles this complex issue well. The algorithm helps users produce defensible results that satisfy the requirements of reviewers and colleagues.

While the Pro Group algorithm does protein grouping automatically, the user is encourage to read this document carefully to understand the protein grouping issue. With greater understanding, users are better able to inspect and critically review results.

[1] Bradshaw, R. A., Burlingame, A.L., Carr, S., and Aebersold, R. 2006. Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* 5, 787-788. The guidelines are available at:
http://www.mcponline.org/site/misc/PhialdelphiaGuidelinesFINALDRAFT.pdf

## Detected Proteins – The New Convention in Reporting

Key new language from the second version of the MCP guidelines describes a subtle but important shift in perspective for interpreting protein search results:

> *"While the identification of shared peptides implies that multiple related protein sequences are present, the initial assumption should be that only a single form is being detected."*

Historically, there has been a pervasive failure among researchers to grasp this concept and, because it is human nature to want to report the highest number of identifications possible, many proteomic publications reported inappropriately large numbers of identified proteins in the past.

Protein identification should be reframed as protein detection. When viewed this way, the guideline for determining which number of proteins should be reported is much clearer: it should be the number of distinct protein species that can be proven as detected. Where there is a suggestion of a group of multiple related proteins, the default assumption must be that only one form is detected until proven otherwise. Evidence used to prove the detection of one protein cannot be used again to prove the detection of a second protein. This philosophy is at the core of the Pro Group algorithm.

For each detected protein, there might be ambiguity as to exactly which protein sequence is being detected. The Pro Group algorithm bundles each detected protein with related redundant proteins into a protein group. Redundant proteins should be shown to make the ambiguity clear, but they should not be used to increase the number of proteins reported.

The protein identification results in the ProteinPilot software are designed to show the results of the Pro Group algorithm. A list of detected proteins is shown and, for each detected protein, the related proteins.

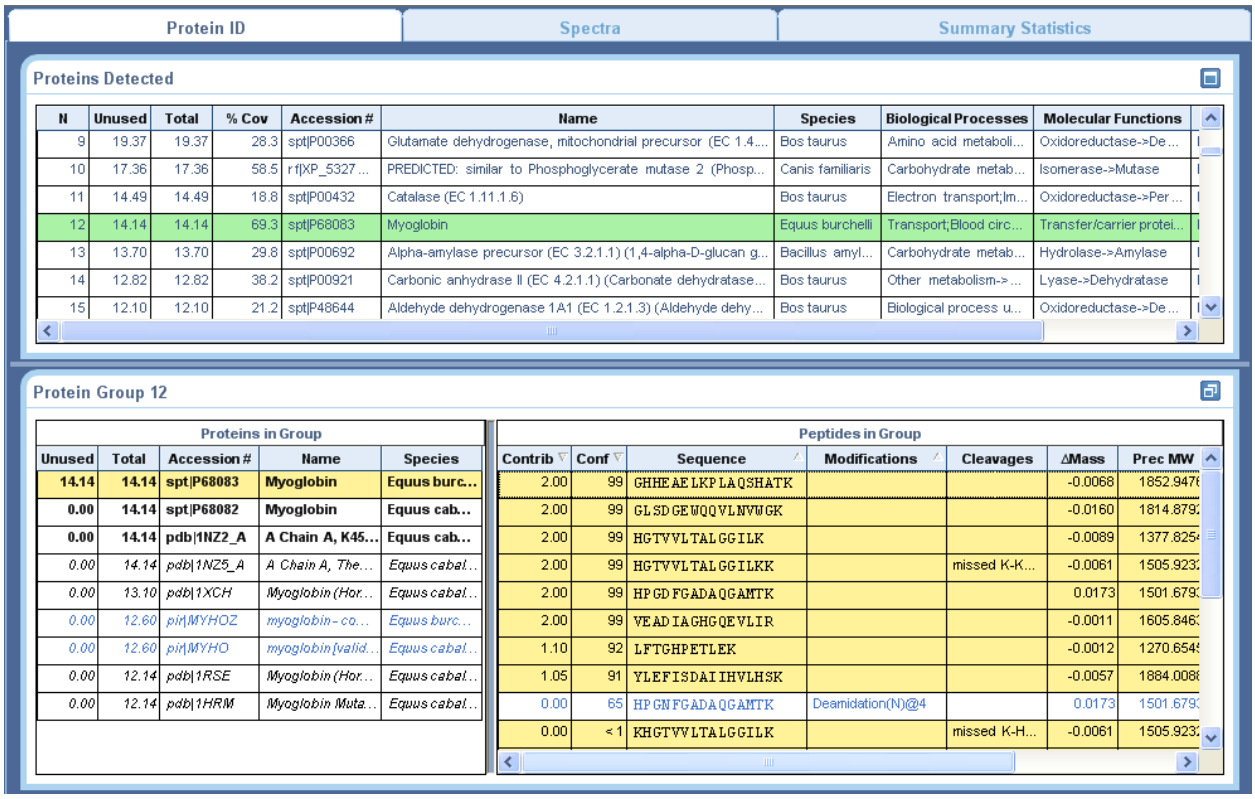


**Figure 1 – The Proteins Detected table (top) and the Protein Group pane (bottom) on the Protein ID tab of ProteinPilot software**

The **Proteins Detected** table contains the list of proteins believed to have been detected. The number of proteins to report is either the number of proteins in this table or, preferably, this number can be adjusted based on the false discovery rate (FDR) analysis results. Regardless of which method used to determine the minimum **Unused ProtScore** threshold, no protein in this list is justified on the basis of evidence already claimed by a higher-ranked protein. For each protein detected, the ProteinPilot software shows the group of related proteins in a separate pane – the **Protein Group** pane. In Figure 1, the twelfth most-confident protein detected, myoglobin, is selected, to see the details of its group.

The display of a protein group has two simple goals:

- Indicate ambiguity as to which protein sequence is actually being detected.
- Indicate where more than one related protein has been detected.

The **Proteins Detected** table selects one protein from the group as the representative winner protein. The **Protein Group** pane shows the degree of certainty that this specific sequence from a database is being detected. Different text colors and fonts are used to denote the relationship of each protein in a group to the representative winner protein.

The following sections explain the different types of protein relationships in a group. For each protein relationship, the document explains why the protein is worth seeing in results, even if it should not be used to increase the number of proteins reported as detected.

# Basic Relationships of Proteins in a Group

This section explains the simpler relationships within a group and how these relationships are shown in the ProteinPilot software.

## Equivalent Winner Proteins and Subset Proteins

The two simplest relationships between a protein in a group and the winner protein are:

- Proteins with exactly the same set of identified peptides belonging to the winner. These are **equivalent winner proteins**.
- Proteins with only a subset of the peptides identified as belonging to the winner, and nothing more. These are **subset proteins**.

In Figure 2, the fifteenth ranked protein is selected in the **Proteins Detected** table (in green). The details for the protein group relationships for this protein are shown in the **Protein Group 15** pane.

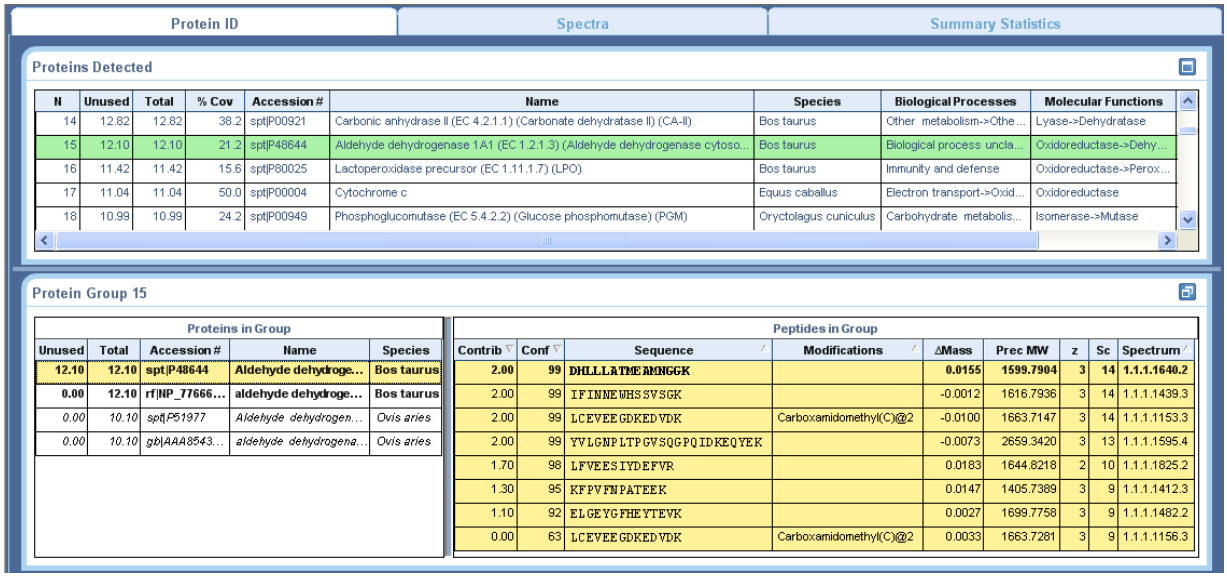


**Figure 2 – An example group with multiple equivalent winners and subset proteins**

One protein identifier, **spt|P48644**, is listed as a representative of the group in the **Proteins Detected** table at the top. In the **Proteins in Group** table, four protein sequences are considered relevant by the Pro Group algorithm:

- Two proteins (shown in bold black text) are equivalent winner proteins. They are considered equivalent because they explain exactly the same set of peptides. They are winners because they have the most evidence of any protein in this group. The first one acts as the arbitrarily chosen representative of the group in the **Proteins Detected** table.
- Two proteins (shown in black italic text) have a subset of the peptides explained by the winners in the group, and nothing more. They are subset proteins.

## Using Venn Diagrams to Explain Protein Groups

These relationships are easily illustrated using Venn diagrams to indicate which proteins have peptides explaining which spectra. Figure 3 shows a Venn diagram that corresponds to Figure 2.

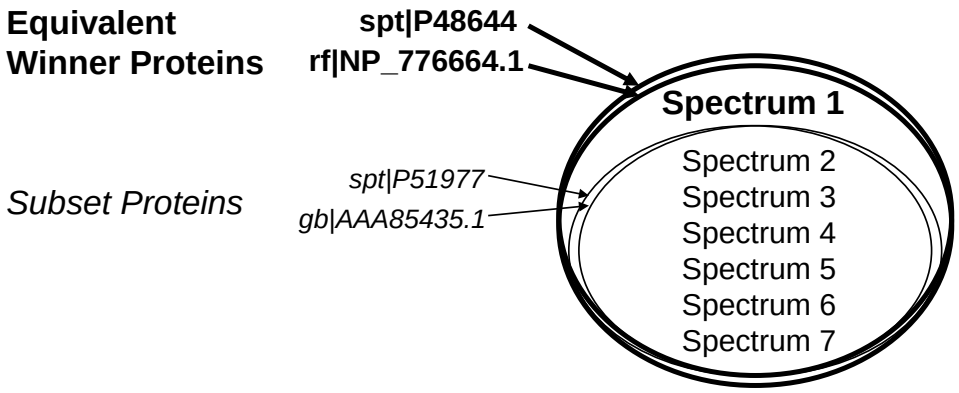


**Figure 3 – Venn diagram of spectra explained by equivalent winner proteins and subset proteins in a group**

Each circle represents a protein sequence in the database, indicated here by the accession numbers pointing to each ring. The spectra inside any ring are the spectra that are explained by these proteins because the protein has a peptide sequence that is a viable identification for the spectrum. These diagrams are useful to explain grouping relationships and are used extensively in this document, so it is important to understand this representation before proceeding.

The formatting in the Venn diagrams is similar to the formatting in the ProteinPilot software user interface. For example, equivalent winner proteins are shown in bold black text in the software and with bold black rings in the Venn diagrams. The subset proteins shown in non-bold italic text in the results are shown with thinner black rings in the diagrams. Lastly, bold black formatting of peptides in the **Peptides in Group** table indicates that a sequence belongs only to the winners

presented in the group. In Figure 3, the spectrum from which a peptide was identified that is specific to the winners is also shown in bold black.

Because the peptides unique to the winner are shown in bold black text, it is clear that this peptide is the critical one to inspect to confirm the selection of the equivalent winner proteins over the subset proteins. For simple groups like this, it is easy to understand the relationships in a protein group from the formatting. For more complicated relationships, use the selection tools in the software.

For this example, if a user clicks one of the equivalent winners and then control-clicks one of the subset proteins, the display indicated in Figure 4 is shown.

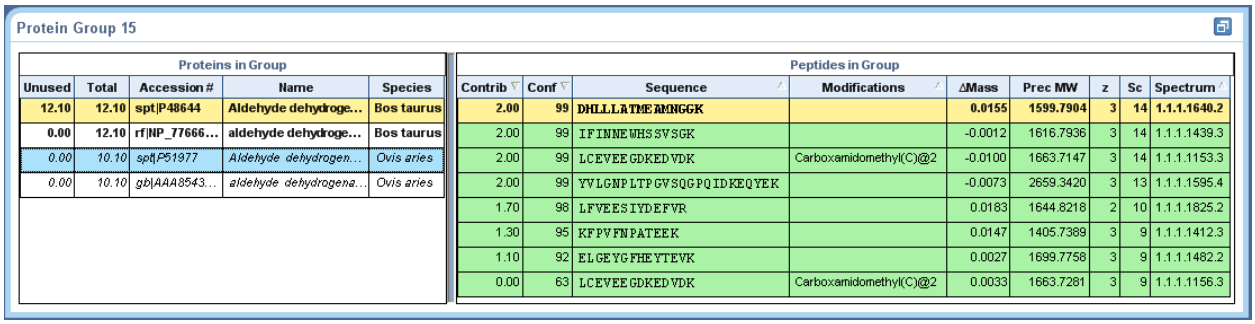


**Figure 4 – Protein-protein intersection in the Protein Group pane**

The first selected protein is highlighted in yellow and the second in blue. All peptide sequences specific to the yellow protein are shown in yellow and all peptides specific to the blue protein are blue (none in this case). All sequences common to both selected proteins are highlighted in green (all but one in this case). This conveys the same information as the Venn diagram in Figure 3.

## Competitor Proteins

In the previous section, subset proteins were shown in a group, but are all subset proteins worth showing in a result?

Note that the winners in the previous example differ from the subset proteins by only one peptide identification. If this one peptide (shown in bold black) were to be an incorrect identification, then the two subset proteins would be as good an explanation for the data as the reported winners. Because these subset proteins are close to being as correct as the equivalent winners shown, they are considered to be competitor proteins. They are close enough to being correct that they should be kept in view.

There are likely to be many other subset proteins that explain a much smaller number of the spectra in the group, but they are not considered competitor proteins by the Pro Group algorithm. It would require that a large number of the winner's peptide identifications be wrong for these proteins to actually be as good an answer as the winners. The ProteinPilot software does not show these uncompetitive proteins in a protein group.

Some software might show all subset proteins or only those that have a certain total score. By showing only competitor proteins, the Pro Group algorithm lets users focus on a smaller more relevant set of subset proteins, namely the proteins that actually have a chance of being the best explanation for a particular portion of the data.

## Spectra Are the Evidence, Not Sequences

Proteins are compared based on the spectra they explain, not the identified peptides they contain.

This concept is one of the advantages of the Pro Group algorithm. Comparisons of this type prevent falsely reporting redundant proteins and still allow tracking of competitors. The example in Figure 5 illustrates this concept.

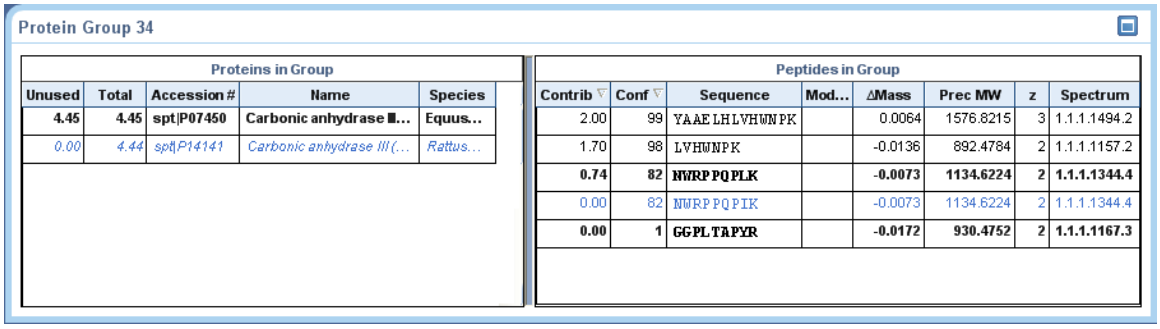


**Figure 5 – A Protein Group including a protein that has a peptide not belonging to the winner**

Here there are two similar forms of carbonic anhydrase, where the one in bold black text is the winner protein, and the second protein is shown in blue italic text. Blue text indicates evidence for a peptide not found in the winner of the group. A peptide is shown in blue if its sequence is not found in the winner, and a protein is shown in blue if it has at least one peptide whose sequence is not found in the winner.

Clicking the first protein and then control-clicking the second protein shows the display indicated in Figure 6.

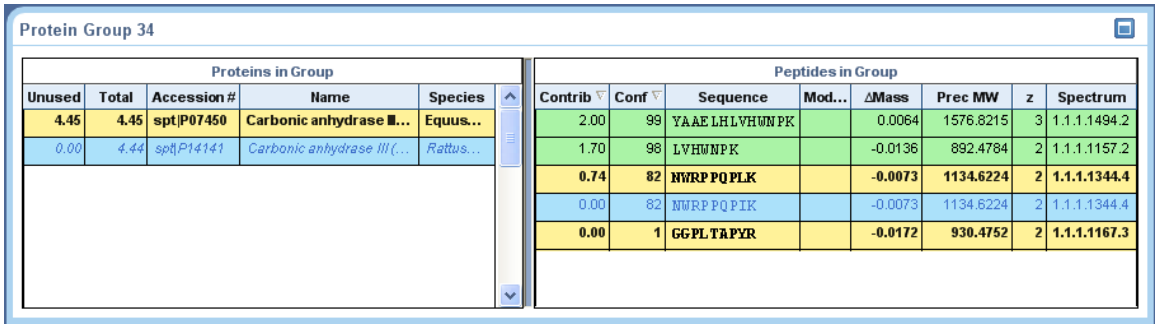


**Figure 6 – Protein-protein intersection when not all peptides belong to the winner**

Now it is clear that the blue peptide is specific to the blue protein. There are also two peptides shown in bold black because they are specific to the winner protein.

Given that each of these two proteins has at least one identified peptide distinct from the other, it might seem that these proteins are not redundant. Users might think that it would be valid to report both in a list of detected proteins, but this is not true.

Looking at the **Spectrum** column in the **Peptides in Group** table, the spectrum index is the same for the blue peptide and one of the bold black peptides. This means these two peptides are alternate hypotheses to explain the same fragmentation spectrum (from cycle 1344, experiment 4). The only difference between these two peptides is that one has isoleucine where the other has leucine. While the two peptide sequences are distinct, the spectral evidence for these peptides is the same.

In Figure 7, the Venn diagram on the left is a spectrum-centric view and the one on the right is a sequence-centric view.

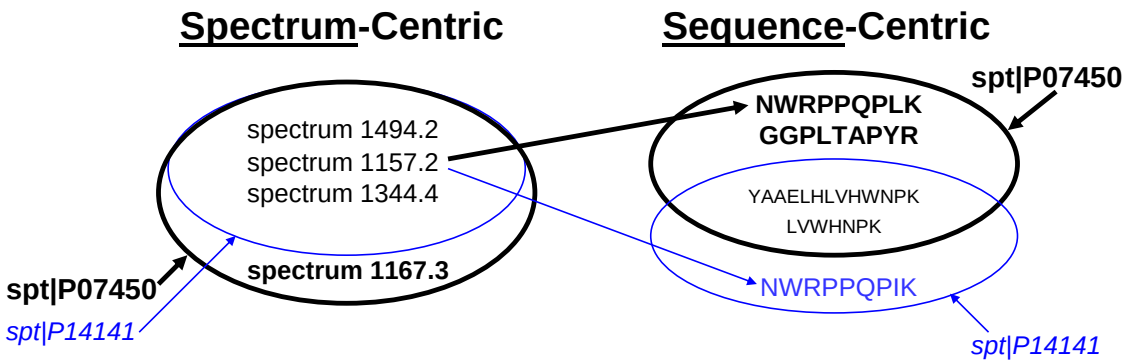


**Figure 7 – Spectrum-centric versus sequence-centric Venn diagrams of the evidence for two proteins**

One of the fundamental concepts of the Pro Group algorithm is that it is evidence-centric: the goal of protein grouping is to find the smallest number of proteins needed to explain all the fragmentation spectral evidence. Thus, what really matters in the figure above is the spectrum-centric view on the left. Only one protein is needed to explain all these spectra. Some other software programs would report both of these proteins, because both are needed to explain all the peptides. For example, if the **Require bold red peptide** option is not selected in Mascot, both of these proteins are included.

However, excluding this protein is not appropriate either, as both of these proteins are relevant. If the one marginal identification from spectrum 1167.3 is incorrect, then the blue protein could be considered the winner. This blue protein is a competitor protein, even though it is not a subset protein. Thus, you want to keep this protein in view with the winner. It would be sufficient to review the data and decide to report the blue protein instead of the winner, but it should not be reported as a detected protein in addition to the winner. Counting both proteins in

a case like this is a common example of how insufficient protein grouping leads to an inflated number of proteins reported. With other software, even if reporting incorrect additional proteins can be prevented in a situation like this, the connection between a winner protein and a relevant competitor like this blue protein will be lost entirely or very difficult to discover. The Pro Group algorithm and its method of presenting results make it easy to see relevant competitor proteins like this.

## Grouping Does Not Consider Unobserved Sequences

It is important to note that protein grouping in the Pro Group algorithm only considers portions of sequences in proteins for which there is observed evidence, rather than the entire sequence of the proteins. This is in contrast to other sequence analysis software, such as a BLAST alignment. The reason for this difference is that the purpose is different. A BLAST analysis assesses the similarity between sequences in a database, independent of experimental data about these sequences. Protein grouping, on the other hand, tries to assess which proteins have been detected based on experimental observations. Unobserved regions of protein sequence, by definition, play no role in explaining the data.

Figure 8 shows three different protein accession numbers that are considered equivalent winners for the detection of one protein.

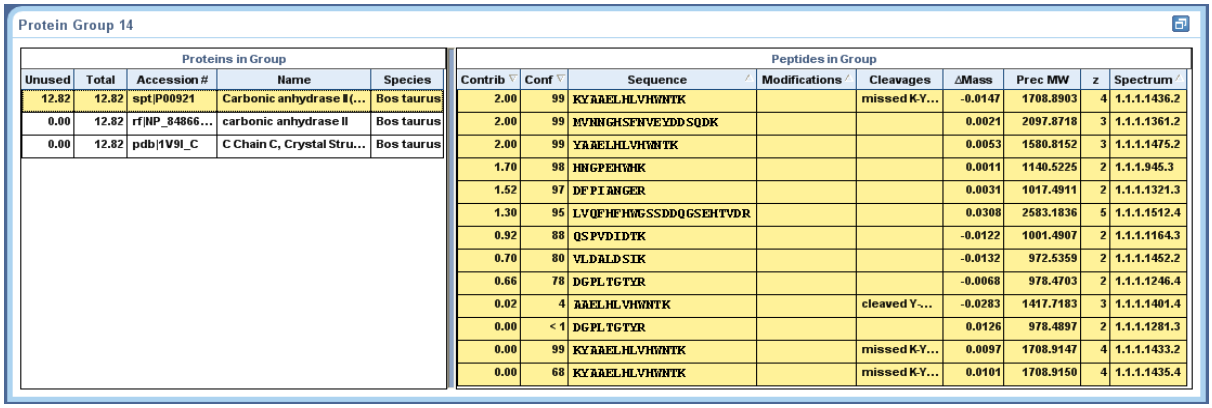


**Figure 8 – A protein group with three equivalent winner proteins**

Although each of these proteins has all of the identified peptides in the **Peptides in Group** table on the right, the complete sequences of the proteins might be different. Click a row in the **Proteins in Group** table to show the complete sequence for that protein in the **Protein Sequence Coverage** pane, as shown below for the protein **spt|P00921**. The areas in grey represent portions of the sequence with no spectral evidence. Identified sections of the sequence are color coded by confidence. Low confidence peptides are red, moderate confidence peptides are yellow and high confidence peptides are green.

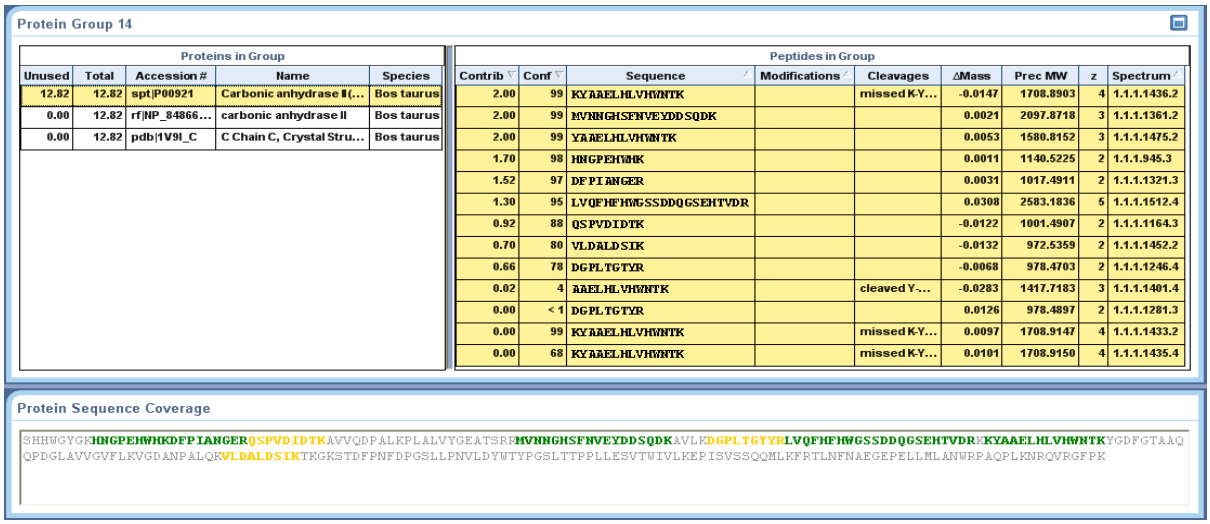


**Figure 9 – Protein Sequence Coverage display, with grey areas indicating missing spectral evidence**

The sequence is not completely covered by the peptides that have been observed. The protein sequences for the three proteins show that each protein actually has a slightly different sequence at the N-terminus.

Figure 10 shows two views of this situation. On the left is the comparison of the complete sequences, where we can see the first tryptic peptide is distinct in each protein. On the right is the Pro Group algorithm's view of the three proteins, with only the peptide subsequences for which there is experimental data.

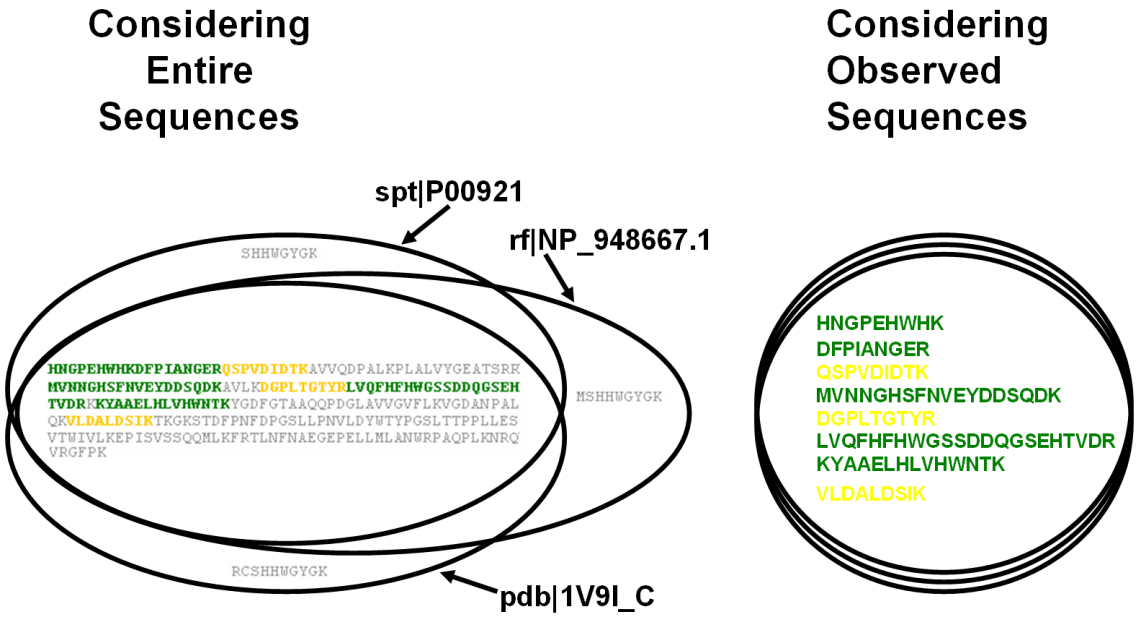


**Figure 10 – Venn diagrams considering entire sequences versus observed sequences**

To determine which of these three protein sequences is being detected, users could acquire additional mass spectral data. Specifically, the three theoretical peptides that differentiate between these forms could be targeted by using an inclusion list during data acquisition.

## Detection of Multiple Related Protein Forms

So far, this document has described the redundant proteins that are included in a protein group with the winner protein. There are cases, however, when two or more proteins share much of the same spectral evidence, but are not redundant. There might still be enough evidence to justify declaring multiple similar proteins as detected, without using the same data twice to justify multiple proteins. A protein group in which more than one protein can be declared detected is called a multi-detection group. This section explains how the Pro Group algorithm determines when a protein group is a multi-detection group.

Below is an example of a multi-detection group. Figure 11 shows the ninth protein in the **Proteins Detected** table. This is the highest ranked form of glutamate dehydrogenase in the table.

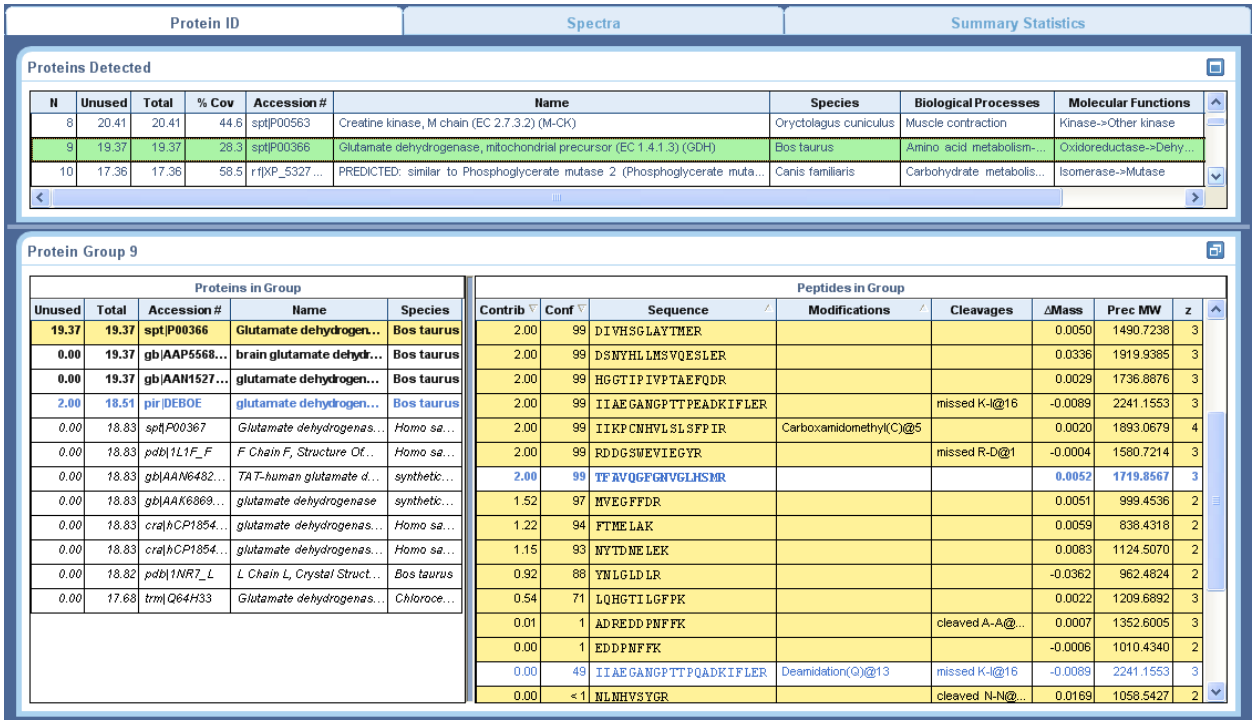


**Figure 11 – A multi-detection group for glutamate dehydrogenase**

Note the different formatting here: bold blue proteins and bold blue peptides. Blue formatting still has the same significance, indicating proteins or peptides with sequence that is distinct from the winner (bold black) protein. Bold blue formatting is used for proteins or peptides that are particularly important.

Proteins are shown in bold blue if they claim enough evidence to be declared detected. Proteins in bold blue text indicate a multi-detection group.

Peptides are shown in bold blue if their sequences are distinct from the winner protein and if they can contribute to the detection of another protein in addition to the winner. A peptide can contribute to the detection of an additional protein form if:

- it is identified with some confidence, and
- it is identified from spectral evidence not already used by the winner protein.

In a multi-detection group, the highest ranked protein form is called the primary protein form. It is highest ranked because it is the one that explains the most spectral evidence of all the proteins in the group. If only one form is being declared as detected, the primary form is the most obvious choice. The other related detected forms are called secondary protein forms. In Figure 11, the winner is the primary form. It explains a lot of spectral evidence, so it is a high confidence detection that is ranked ninth in the list of proteins.

Because only one entry is shown in the **Proteins Detected** table for each detected protein, where bold blue proteins are shown, this indicates that these proteins appear elsewhere in the **Proteins Detected** table. Scroll down the table to see that the glutamate dehydrogenase in bold blue in the group for the primary form (shown in Figure 11) is the 46$^{th}$ protein in the **Proteins Detected** table as shown in Figure 12.

The **Protein Group** pane is designed to explain the detection of one detected protein. To explain the proteins in a multi-detection group, the same group is shown multiple times, once for each detected protein. If the secondary form of glutamate dehydrogenase is clicked in the **Proteins Detected** table, another instance of the same protein group is shown, but with the secondary form shown as the winner.

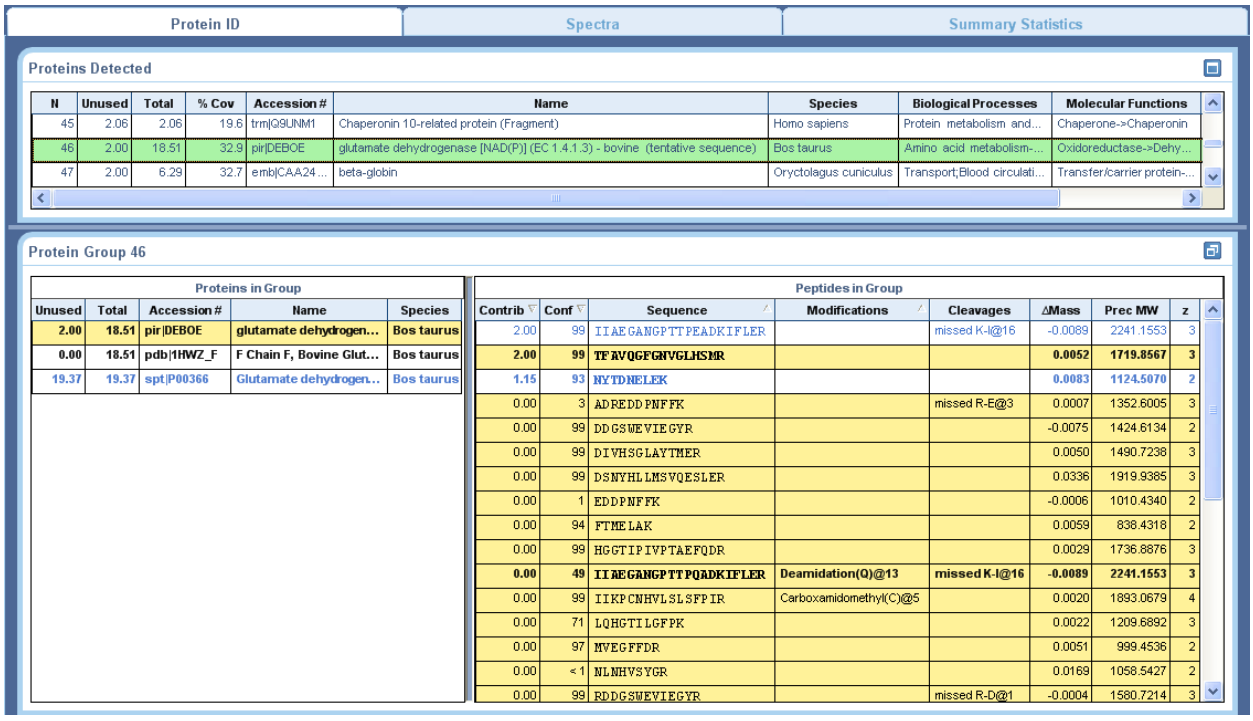


**Figure 12 – Instance of a multi-detection group explaining the secondary form**

The same peptides are shown but they are formatted differently. The proteins shown in each case are similar but not identical. In each instance of the group, the formatting conveys information about the winner for that group. Figure 13 shows a simplified picture of how each instance of this group is rendered.

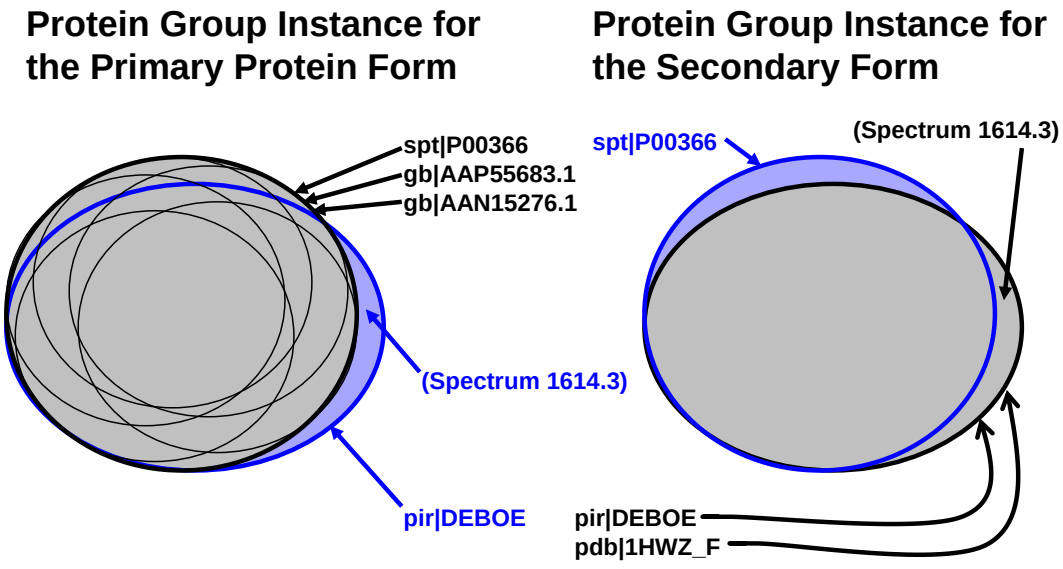


**Figure 13 – Venn diagrams for protein and peptide display in different instances of a multi-detection group**

The Venn diagrams parallel the formatting used in the software. For example, heavy black rings mimic the bold text used for detected proteins and light rings mimic the non-bold italic text used for related proteins that are not detected. Exactly equivalent proteins are shown as a single ring for simplicity. Although not all of the spectra explained in each Venn region is listed here, these are spectrum-centric figures.

The diagram on the left shows how the group is displayed when the primary form is reported. All equivalent winners and all competitive subset proteins are shown in the table to indicate the ambiguity about which accession is actually being detected. One representative of the detected secondary protein form (shown in bold blue) is listed to indicate that another form is also detected. The diagram on the right shows how the same group of related proteins is displayed for reporting the detection of the secondary form. All equivalent winners of the secondary form are listed, and it has no competitive subsets or other competitors. One representative of the primary protein form is shown (in bold blue) so users can tell how the secondary protein is related to it.

The peptide from spectrum 1614.3 is shown differently in the two instances of the group:

- For protein #9, it is bold blue (and in blue fill in the Venn diagram on the left) because it is the critical distinct evidence supporting another protein form.
- For protein #46, it is bold black because its sequence belongs only to the winner of that group.

## Ranking of Proteins is Based on Unused Evidence

Perhaps the most important feature of the Pro Group algorithm is how it ranks proteins. Some other protein identification software ranks its list of reported proteins based on the total score for the protein. That is to say, all peptide evidence for any protein is counted to determine the rank of the protein. The Pro Group algorithm follows the convention that each new reported protein must be detected based on evidence not already explained by higher ranked proteins. In this context, ranking by total protein score is not the correct way to rank proteins from most likely to least likely. The Pro Group algorithm does not do this, and here is why.

Consider the situation in the previous section, where there is a primary form with a large amount of evidence, and a secondary form with only a single spectrum not already explained (i.e. “unused”) by the primary form. To revisit the same figure, we could indicate the accounting of spectral evidence as shown in Figure 14:

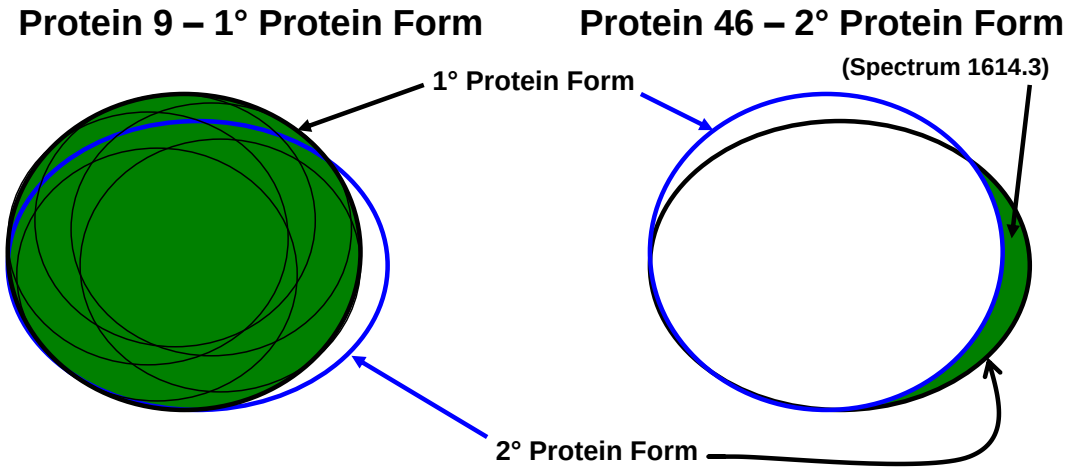


**Figure 14 – Venn diagrams for unused evidence of detected proteins in a multi-detection group**

The secondary form has almost as much total evidence as the primary form. But after the primary form is reported as detected, the only unused evidence is one spectrum, corresponding to the area shaded green in the diagram on the right. If the identification for this spectrum is wrong, there is no evidence to claim that this protein has been detected in addition to the primary form. Thus, this protein should be ranked among other proteins that are based on single identifications, not among proteins that have the same large total score as this secondary form.

The **Total** and **Unused** columns in Figure 11 and Figure 12 show the measures of the total and unused protein scores, referred to as **Total ProtScore** and **Unused ProtScore** in their unabbreviated forms. The **Unused ProtScore** for the primary form, 19.37, is much higher than that of the secondary form, 2.0, yet the **Total ProtScores** are almost the same, 19.37 and 18.51, respectively. The ranking in the **Proteins Detected** table is based on the **Unused**, not the **Total**, column. This is why the primary form is ranked in the top 10 most likely detected proteins, and the secondary form is ranked much lower among other proteins with only single hits supporting them. If the goal is ranking proteins by decreasing confidence, this makes perfect sense.

To show the impact of this difference, consider the following example. The data set was searched using the Paragon™ algorithm in **Rapid** mode and with Mascot, with the modifications set so that the search space was essentially the same in both cases. To reduce the Mascot results to a more appropriate number of proteins, a bold red peptide was required and the minimum ions score was set to be equal to the significance score. Essentially the same number of proteins was

found in each search, but the way the proteins were ranked is substantially different. Figure 15 and Figure 16 focus on these differences in ranking.

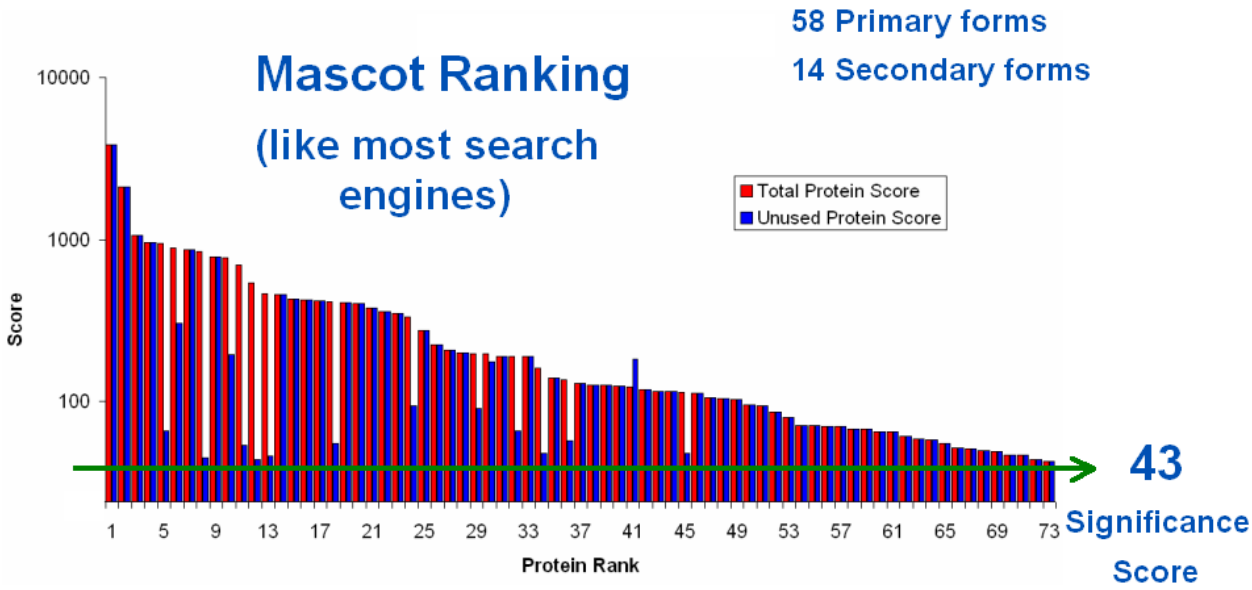


**Figure 15 – Total and unused protein scores for protein ranked by Mascot's total score**

Figure 15 shows the total protein score Mascot reports in red and the amount of unused peptide evidence calculated manually in blue. Any protein where the two bars differ in height is a secondary protein form. Because Mascot ranks proteins by total score, the secondary forms stand out as proteins with lower blue bars.

**NOTE**: Since the original version of this document was written, Mascot has added the option of grouping proteins into protein families, which improves this situation.

There are several very highly ranked proteins, even in the top 10, with unused evidence barely above the significance threshold. This means that there is a very real chance that these proteins have been incorrectly identified because they probably hinged on one peptide identification only.

Now consider the way the Pro Group algorithm ranks its search results for the same file.

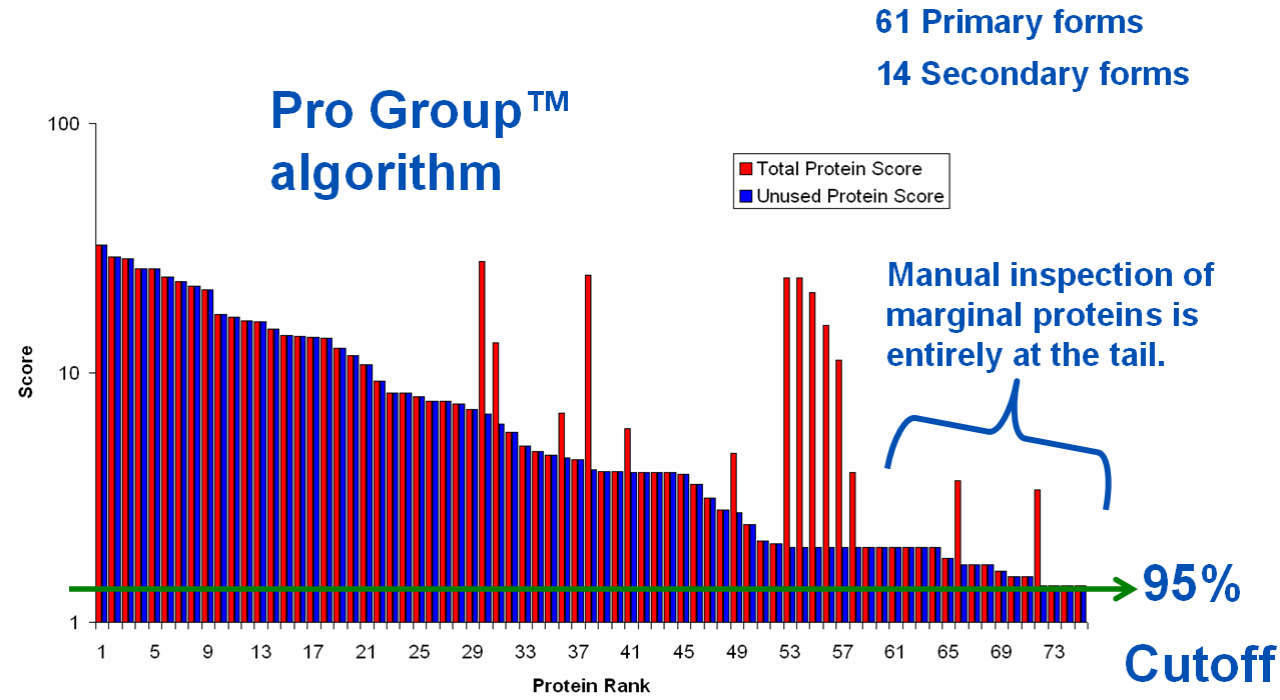


**Figure 16 – Total and unused protein scores for protein ranked by the Pro Group™ algorithm's Unused ProtScore**

In Figure 16, the proteins are ranked by **Unused ProtScore**. The secondary proteins appear as red spikes (**Total ProtScore** > **Unused ProtScore**). The secondary proteins are generally towards the tail end of the results. Because this ordering is the real order of decreasing confidence, when results are inspected manually, users can focus entirely on the tail end. For results from other search engines that do not rank this way, the entire list of proteins must be inspected, because single-hit proteins can appear anywhere in the list, as shown in Figure 15.

## Competitor Proteins in Multi-Detection Groups

The concept of competitor proteins can be explored in more detail by considering the Venn diagrams in Figure 17.

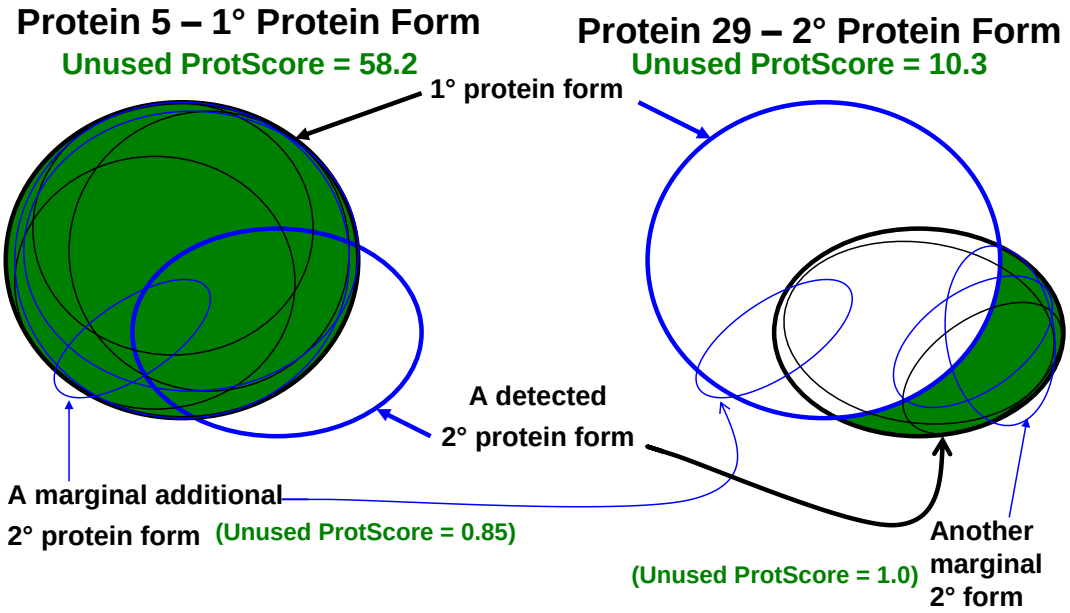


**Figure 17 – Venn diagrams for a multi-detection group including marginal secondary forms**

Although previous sections did not state this explicitly, the area of each region in these Venn diagrams is intended to be proportional to the quantity of evidence for that region. In this example, we have a group of related proteins where there are two different forms that are detected above a confidence threshold of 95% (**Unused ProtScore** = 1.3). The primary form explains the vast majority of the spectra, indicated by the green area in the diagram on the left. The detected secondary form also explains several additional spectra, making a solid case for its presence. There are also two additional marginal secondary forms without sufficient evidence to be declared. They are called marginal because they do have some small amount of unused evidence, but not enough to exceed the **Detected Protein Threshold** of 1.3 (corresponding to 95% confidence).

The goal of reporting competitor proteins is to list all protein accession numbers that are close enough to explaining the same set of evidence as the reported winners that they might be the true answer. In a multi-detection group, each detected protein has a different set of competitors. In Figure 17, the competitors of the primary protein form on the left are the black subset proteins and the blue italic proteins that have a subset of the same spectra, although not the same exact sequences. One or two incorrect peptide identifications specific to the primary form might mean that one of these competitor proteins is actually the best answer. The secondary protein is not a viable competitor to be the right answer instead of the primary protein form, but it is shown because it is a related additional detected

form. The marginal secondary forms shown in this figure would not be shown in the instance of this group for the primary form, unless the **Detected Protein Threshold** was lowered.

For the secondary form, the definition of a competitor must be clarified. In the diagram on the right, the unused evidence newly explained by reporting this protein form is the area in green. Thus, a competitor protein is any protein that explains all (or nearly all) of this green area, the unused evidence. It is not necessary for a competitor protein to explain nearly all of a secondary form's total evidence. Thus, all of the proteins shown in this diagram, except the primary form and the marginal secondary form on the bottom left, are viable competitors to the secondary form.

## Competitor Proteins are Important

One of the advantages of the Pro Group algorithm is the ability to view competitor proteins for each group. Understanding which proteins are relevant competitors is challenging and it is becoming more and more apparent how important it is to find competitor proteins.

The second 'Paris' version of the MCP guidelines stated:

> *"The apparent ambiguity in peptide assignment requires reporting of a protein group."*
>
> *"Authors should explain and be able to justify cases where a single protein from a protein group has been singled out or that more than one member of a protein group is present."*

This suggests that when the selection of one form from a protein group cannot be justified, then ambiguity among several forms should be reported. Most reviewers would probably accept reporting only the equivalent winners when reporting this ambiguity. However, there can easily be proteins that explain the same set of spectra but differ in sequence in only one spectrum. For example, two proteins might differ by only isoleucine versus leucine in a single position. Two proteins with only this difference are exactly equal and should be reported as part of the ambiguity among accession numbers, but some other software does not do this. Other software is typically only able to recognize and group proteins that have evidence for exactly equal sequences. In contrast, the Pro Group algorithm can group these types of proteins and can also show inexact competitors, allowing users to keep all viable alternatives in consideration.

Tracking relevant competitors is not only valuable for publication requirements; it can be critical to research. This is particularly true when trying to compare your own results or compare your results to those from a publication or data repository. Suppose one set of MS/MS data is acquired on a sample, and the first list of proteins detected is determined. Any of the following could be done next:

- Acquire additional data on the same sample using inclusion/exclusion lists to “dig deeper”.
- Acquire data on a different instrument or by a different ionization technique.
- Repeat the exact same run.
- Compare your results to those of someone else.

Each of these sources of additional data will sample a slightly different set of peptides. The best protein accession number from a search might change, even though the same protein species is being detected. Reporting only the top scoring accession number would make it very hard to recognize that the same protein is detected in both cases. However, a group is reported that includes viable competitors with both exact and inexact relationships instead, there is a much greater chance of recognizing that the same protein has been found.

## Summary

The protein grouping issue is something that should be understood to interpret results from bottom-up proteomics experiments. The Pro Group algorithm and the display of its results in ProteinPilot Software provide a solution to this issue. Here is a review of the key questions this document answered.

- What is the protein grouping issue?

The protein grouping issue can be stated as follows:

> *Given peptide identifications from spectra acquired in a bottom-up proteomics experiment, determine the list of non-redundant proteins that can justifiably be declared as detected. Where there is ambiguity about which specific database sequence is best to declare detected, present the competing alternatives.*

It is described as a grouping issue because the key to its solution is to find groups of proteins that derive evidence from the same spectra, and to declare as detected only those proteins in the group required to explain significant evidence.

- What kinds of false proteins are reported by software that does not do proper protein grouping?

Without proper protein grouping, many proteins are reported based on spectra already used to justify the detection of more confident protein hits. The most common cause of false and redundant hits is when multiple, slightly different peptides from the same spectrum each are used to justify the detection of multiple proteins. The spectrum might provide proof that one of those peptides is in the sample, but it does not provide proof that all of the similar peptides are in the sample, so that piece of data should not get used multiple times to justify multiple proteins. Another common cause of false proteins is the reporting of proteins with only an insignificant amount of evidence not already used to justify more confident proteins. There are additional more subtle causes of false proteins as well.

- How does the Pro Group algorithm prevent the reporting of these false or suspect proteins?

The Pro Group algorithm only reports proteins based on evidence not already used to justify more confident proteins. The tracking of evidence used is based on spectra, not on identical peptides. Neither a single peptide nor similar peptides from a single spectrum can be used multiple times to justify multiple proteins.

- What is a protein group? What are “competitor” proteins, and how does the ProteinPilot software show them?

A protein group is a group of proteins that derive significant evidence from a shared set of spectra. A protein in a group is only reported as detected if it best explains significant *unused* evidence in this set of spectra – evidence that would otherwise go unexplained. A competitor protein is one that explains nearly the same unused evidence explained by the representative winner. The ProteinPilot software shows a representative winner protein sequence from a group of competitor proteins and also shows all member of the group that explain the same spectra equally or nearly as well. The representative winner protein is shown as the first protein in bold black text, and all other competitor proteins are shown with various formats to indicate their relationship to this representative of the protein detection event.

- I was looking for a particular protein. Why didn’t I see it in the results?

Your protein of interest might have a high total score, but that score might have come entirely or almost entirely from spectra used to justify the detection of an even better protein. If your protein is not even close to explaining the data explained by the detected protein, there is no basis to claim that your protein was detected in addition to the protein reported as detected. There is not even a basis to say that your protein is a competitor protein that might be in your sample instead of the protein reported. If you believe that particular protein is in your sample, you should acquire data in a more targeted way to get evidence specific to that protein.

- How can I tell when other protein identification software is reporting an invalid number of proteins because of the failure to do proper protein grouping?

Unless other software demonstrates an awareness of these issues, you should assume that the number of proteins it reports can be significantly inflated. Here is a list of questions to ask about other software to see if it can produce a defensible number of proteins:

- Does the software require a sufficient amount of unused evidence in order to report a protein?
- Does the software make it easy to find proteins with a borderline amount of unused evidence (even when the total evidence for the protein is high) so that you can manually verify the protein? If the proteins likely to be false are

hidden in the results, you will never know that your number of proteins is inflated unless you manually review *every* protein.

- Does the software track unused evidence by spectrum, rather than by peptide? In other words, does the software forbid using the same spectrum multiple times to report multiple proteins, or rigorously justify any cases where a single spectrum is used more than once?

If the answer to any of these questions is No, then the number of proteins reported by that software can be significantly inflated. Users would need to review their results thoroughly with the protein grouping issue in mind to produce a list of proteins that is scientifically valid and acceptable for publication.

**NOTE**: If users want to review an article that attempts to develop some standardized terminology around many of the concepts in this document, refer to A standardized framing for reporting protein identifications in mzIdentML 1.2. Sean L. Seymour, Terry Farrah, Pierre-Alain Binz, Robert J. Chalkley, John S. Cottrell, Brian C. Searle, David L. Tabb, Juan Antonio Vizcaíno, Gorka Prieto, Julian Uszkoreit, Martin Eisenacher, Salvador Martínez-Bartolomé, Fawaz Ghali and Andrew R. Jones. 4 AUG 2014 DOI: 10.1002/pmic.201400080 http://onlinelibrary.wiley.com/doi/10.1002/pmic.201400080/abstract.

# Revision History

| Revision | Description of Change | Date |
|---|---|---|
| A | First release | March 2009 |
| B | Content updates related to ProteinPilot software v. 4.0 | September 2010 |
| C | Content updates related to ProteinPilot software v. 5.0 | September 2014 |