ProteinPilot™ Software 2.0 for Protein Identification and Expression Analysis

Providing expert results for non-experts and experts alike

ProteinPilot™ Software Overview

ProteinPilot Software transforms protein identification and relative protein expression analysis for discovery research. The software combines the industry standard Mascot search engine from Matrix Science and the revolutionary new Paragon™ Algorithm from Applied Biosystems with the industry-leading Pro Group™ Algorithm. The Paragon Algorithm is a unique new technology that enables simultaneous searching for hundreds of biological and other modifications, genetic variants and unexpected cleavages without the typical explosion of false positives that plagues traditional algorithms. The Paragon Algorithm’s underlying methodology also allows for a unique-friendly user interface for protein identification that is vastly simplified over traditional algorithms.

After the initial search, the Pro Group Algorithm assembles the peptide evidence from the Paragon Algorithm into a comprehensive summary of the proteins in the sample. The Pro Group Algorithm addresses the protein grouping problem by correctly handling the complexities posed by protein subsets and isoforms and minimizing the reporting of false positives.

For relative quantitation studies, ProteinPilot Software supports many chemistry strategies, including ICAT® Reagents, iTRAQ™ Reagents and SILAC™ Reagent labeling, and provides sophisticated analysis tools for these workflows such as protein isoform-specific quantitation and the ability to curate quantitation results.

Key Advantages of ProteinPilot™ Software

- New in Version 2.0: support for the 4800 MALDI TOF/TOF™ Analyzer, in addition to improved support for QSTAR® and Q TRAP® systems.
- Paragon Algorithm allows hundreds of modifications and substitutions to be searched for simultaneously using a unique approach involving feature probabilities and a new kind of sequence tag algorithm.
- Unique user interface means that you enter the sample information you know rather than guessing at algorithm settings you should not have to understand
- Pro Group Algorithm determines which protein isoforms have been detected and determines form-specific quantitation in label-based quantitation workflows.
- Support for the new iTRAQ™ 8plex Reagents, ICAT® Reagents, and an expanded list of SILAC™ reagent labeling schemes for protein expression analysis workflows, including the ability to manual curate results.
- Flexibility with the option to choose the Paragon Algorithm or the Mascot search engine for protein identification.

www.appliedbiosystems.com

Figure 1: Overview of ProteinPilot™ Software Workflow.

The software processes LC/MS/MS data (top: example chromatogram), identifies peptides from the MS/MS spectra (middle: example peptide ID) and assembles peptide identifications into a list of reliable protein identifications (bottom: a reported protein group and associated peptides).
The Revolutionary New Paragon™ Algorithm

The Paragon Algorithm is an innovative new technology that changes the paradigm for protein identification. It employs a feature probability approach that enables the simultaneous searching for hundreds of modifications, substitutions, and unexpected cleavages without the concomitant increase in false positives. This unique feature probability methodology enables a new level of ease-of-use for controlling the search method (Figure 2).

A Radical New User Interface for Search Method Creation

The user interface for the Paragon Algorithm is like nothing else in protein identification software. It simply asks for sample information in biologist’s terms, such as the digest agent and cysteine alkylation reagent used to prepare the sample. You do not have to understand the algorithm and make a series of careful decisions about settings to get good results. You do not have to specify mass tolerances, individual modifications to search for, expected fragment ion types, or exceptions to cleavage rules like missed or semi-specific cleavages. All of these decisions are made automatically based upon the sample treatment and experimental goals.

Identification of Many Modifications Simultaneously

The Paragon Algorithm recognizes that modifications can happen with different frequencies depending upon the sample and how it was prepared. Different probabilities are assigned to different modifications – which is how the algorithm can search for many variants without exploding the time needed for search or the number of false IDs.

For example, when iodoacetic acid is chosen as the alkylating reagent, the Paragon Algorithm assumes carboxymethyl is highly likely on any cysteine residue, and there is a smaller chance that the cysteine will remain unmodified. However, the software also knows that iodoacetic acid leads to iodination of histidine residues as a less frequent side product. If Thorough ID mode is selected, the software automatically searches for these types of less likely modifications.

Special steps in sample preparation or data acquisition greatly influence what modifications users expect to find. Choosing Special Factors tells the algorithm about such details, so it can automatically search for the right modifications to the appropriate degree. For example, selecting Phosphorylation emphasis when a phosphoprotein enrichment step has been employed during sample preparation automatically tells the software to increase the probability of finding phosphorylation modifications on serine, threonine, and tyrosine.

Figure 2: The Paragon™ Algorithm Method Creation Window. The simplified user interface requires only knowledge of the experimental details of the sample, enabling even users new to protein identification to obtain high quality results.

www.appliedbiosystems.com
Industry Leading Pro Group™ Algorithm for High Confidence Protein Detection

Once peptides are identified, the Pro Group™ Algorithm performs a statistical analysis on the peptides found to determine the minimal set of confident, justifiable protein identifications. It obeys a simple principle: You cannot use the same data multiple times to justify the detection of multiple proteins. While seemingly obvious, the failure of commonly used ID tools to obey this rule is a major cause of the false protein identifications that plague proteomics research.

To enforce this principle, the Pro Group Algorithm calculates two ProtScores for each protein. A peptide’s ability to contribute to these ProtScores is based on its confidence, with higher confidence peptides able to contribute more. The Total ProtScore is based on all peptides pointing to a protein, but peptide identifications can only contribute to the Unused ProtScore of a protein to the extent that their spectra have not already been used to justify more confident proteins. Proteins are only reported as detected if they have sufficient Unused ProtScore (Figure 3). The result is the positive determination of proteins actually present in the sample with accurate protein confidences reported.

When redundant proteins and homologous proteins are found, the Pro Group Algorithm organizes related proteins into protein ‘groups’. Multiple protein isoforms in a group are only declared present if unique evidence exists for each isoform.

An Intuitive Display for Fast Comprehension of Results

Figure 4 shows how a protein group is displayed in a Protein Group pane. The proteins in the group are shown on the left, and all the peptides from those proteins are shown on the right. The interactive view allows you to quickly understand the peptide evidence for each protein and the unique evidence justifying the existence of closely related isoforms.

Figure 3: The Protein Grouping Problem. Three proteins sharing peptide evidence are shown. With 8 strong peptides (shown bold in blue and magenta), it is obvious that Protein A should count toward the number of proteins detected. Protein B should not. While its Total ProtScore may be high based on the 4 (magenta) peptides, its Unused ProtScore is 0 (i.e. all evidence has already been used to justify Protein A). Protein C has a peptide unique to it (shown faintly in pink), but its confidence is so low that it is likely false. Adding Protein C to the count of proteins detected based on its high Total ProtScore would be a mistake. The Unused ProtScore is negligible, only coming from this one low confidence ID. The Pro Group’s use of peptide confidences prevents claiming Protein C as detected. The user is not burdened with specifying arbitrary peptide thresholds.

Figure 4: Viewing the Pro Group™ Algorithm Grouping Analysis. Selecting a protein highlights it in yellow, as well as all the peptides that belong to it. Selecting a second protein (by <Ctrl>-click) highlights that protein in blue. Peptides belonging to the first protein and not the second remain yellow. Peptides belonging only to the second protein are highlighted in blue. Peptides shared by both proteins are highlighted in green, allowing instant comprehension of shared versus distinct peptide evidence.
More Information at your Fingertips

ProteinPilot™ Software's innovative user interface allows easy viewing and navigation of results. The results display is divided into three tabs: a Protein ID tab, a Spectra tab, and a Summary Statistics tab. When quantitation is performed, a 4th tab, Protein Quant, is added.

The Protein ID tab shows the proteins detected. The protein-peptide associations for each protein group are also shown (Figure 4). A sequence coverage map, color-coded based on peptide confidence, is available for each protein as an additional way to assess protein confidence (Figure 5).

To see the MS/MS spectrum and the matching sequence ions for the identified peptide sequence, a single click on a peptide opens the raw MS/MS spectrum and highlights the matching peaks (Figure 6). Green cells in the fragment ions table indicate which ions were used to determine the score.

Once a protein has been confidently identified, the next step is to understand the biological significance of that protein or its change in expression. Functional protein classification information, such as molecular function or biological process, is displayed alongside protein ID results when database searches are performed against FASTA files possessing PANTHER annotation.

An overview of the acquisition and results can be viewed on the Summary Statistics tab (Figure 7). An often striking metric is the comparison of the number of identified proteins before and after grouping.

![Protein Sequence Coverage Map](image)

**Figure 5:** Protein Sequence Coverage Map. To further understand the evidence explaining the protein identification and to aid in protein characterization, a sequence coverage map, colored based on peptide confidence, is available. Green indicates the peptide has been identified with at least 95% confidence and yellow indicates at least 50% confidence. Regions identified with lower confidences are shown in red.

![Visualize Peptide Identification Evidence](image)

**Figure 6:** Visualize Peptide Identification Evidence. The matching of the theoretical sequence ions of the identified peptide to the raw MS/MS spectrum is easily viewed in the Fragmentation Evidence window.

![Summary Statistics Table](image)

**Figure 7:** Summary Statistics Table. For a quick overview of the total results for the protein identification experiment, the summary table displays the number of proteins detected, the number of distinct peptides identified for these proteins, and the number of spectra used to identify these peptides.

www.appliedbiosystems.com
Process Protein Expression Data with ProteinPilot™ Software

Protein expression analysis data is both complex and feature-rich. A streamlined analysis requires that powerful analytical tools be coupled with an organized and logical presentation that puts the results you need where you need them. ProteinPilot Software provides a reliable, sophisticated solution to this challenging data analysis problem.

Relative protein quantitation information is reported for each protein in the Protein Quant tab (Figure 8). Peptide ratios are determined and corrected for any systematic bias. Using the results from the Pro Group™ Algorithm, an average ratio is calculated for each protein. P-values are computed for each ratio, providing a convenient measure of its statistical significance. Each measured protein ratio is color-coded based on its P-value, enabling you to quickly focus on the proteins that show real changes in expression.

Other powerful features for quantitation include the following:

- Support for a wide range of labeling schemes, including the new iTRAQ™ 8plex reagents, ICAT™ reagents, an expanded range of SILAC™ reagent labels, as well as protein level schemes for gel work.
- Evidence for peptide quantitation (either raw MS or MS/MS) is easily viewed for selected peptides (Figure 8).
- The Pro Group Algorithm enables isoform-specific quantitation, where differentially regulated proteins with closely related sequences can be accurately distinguished and quantified.
- You can ‘curate’ or determine if a peptide is used in the calculation of a protein’s ratio.
- In a single processing step, both the identification and differential expression information is determined.
- Results can be exported for further analysis with 3rd party tools such as Microsoft Excel.

Figure 8: Visualizing the Quantitation Evidence. The Protein Quant view enables you to easily navigate the protein quantitation information. Color-coded protein ratios highlight differentially expressed proteins. Additionally, a single click displays the MS or MS/MS quantitation evidence for each peptide.
Confirm Identification Results with a Mascot Search

To add flexibility, the software allows users to choose the Mascot® search algorithm as the search engine for protein identification. When Mascot is chosen, the familiar Mascot user interface is shown and results are displayed in the familiar Mascot HTML report format (Figure 9). Results can be compared with Paragon™ Algorithm searches of the same data for the validation of protein identification results.

Figure 9: Mascot Search Results. When searches are performed using Mascot from within ProteinPilot™ Software, the familiar HTML report is shown.

ProteinPilot™ Software Viewer Application Makes Sharing Results Easy

Bottom-up proteomics data is very rich and requires powerful tools such as ProteinPilot Software to reduce its complexity. For sharing experimental results with colleagues or publication reviewers, ProteinPilot software can be installed by anyone and used to open result files, functioning as a free viewer. The viewer application has the same powerful user interface as the full licensed version. Sharing results is easy with the portable group results file format. Download ProteinPilot trial version at:

www.appliedbiosystems.com/ProteinPilot

An expired trial version will still function as a viewer application.

Conclusion

ProteinPilot Software sets a new standard for performing protein identification and biomarker discovery experiments. The sophisticated processing tools enable all users, regardless of experience, to obtain reliable, understandable results. New ProteinPilot Software provides the industry standard Mascot search algorithm as well as the powerful new Paragon Algorithm to identify more from your sample – often doubling the number of spectra yielding peptide identifications. The industry leading Pro Group Algorithm sets the bar for reporting reliable, defensible protein identifications. Coupled with powerful mass spectrometers and labeling chemistries for protein expression analysis from Applied Biosystems|MDS SCIEX, ProteinPilot Software provides a significant advancement in the biomarker discovery field.

Publication 114PB16-02

Applera Corporation is committed to providing the world's leading technology and information for life scientists. Applera Corporation consists of the Applied Biosystems and Celera Genomics businesses. Applied Biosystems|MDS SCIEX is a joint venture between Applera Corporation and MDS Inc.

For Research Use Only. Not for use in diagnostic procedures.

Applied Biosystems, Applera and AB (design) are registered trademarks and iTRAQ is a trademark of Applera Corporation or its subsidiaries in the U.S. and/or certain other countries. MALDI TOF/TOF, Paragon, Pro Group and ProteinPilot are trademarks and Q TRAP and QSTAR are registered trademarks of Applied Biosystems|MDS Sciex. ICAT is a registered trademark of the University of Washington and is exclusively licensed to the Applied Biosystems Group of Applera Corporation. SILAC is a trademark of Invitrogen Corporation.

All other trademarks are the sole property of their respective owners.

Information subject to change without notice.

© 2007 Applera Corporation and MDS Inc. All rights reserved.

www.appliedbiosystems.com