IDPicker Overview

July 23, 2007, David L. Tabb, Vanderbilt University[1]

Introduction

IDPicker is software for protein assembly from raw database identifications. It was created in collaboration between the David Tabb and Bing Zhang laboratories in the Vanderbilt University Medical Center department of Biomedical Informatics. The software is available from the Tabb group website: http://fenchurch.mc.vanderbilt.edu/lab/software.html

Protein assembly from peptide identifications involves several steps. The software first discerns which peptide identifications are trustworthy. It groups together duplicate peptide identifications, and it records the links between peptides and proteins. If the user is compiling a report from multiple runs, IDPicker can apply an appropriate hierarchy to structure the data. The maximal protein list, once assembled, can be pared down by filtering proteins based on the number of peptides matched to each and by parsimony-based rules. The software then reports the protein and peptide information via HTML or another format.

This whitepaper will describe some ways in which IDPicker output can be interpreted. It presumes some knowledge of database search identification. The article introducing MyriMatch (PubMedID 17269722) may provide helpful background on this topic. Elias and Gygi's article on reversed searching (PubMed ID 17327847) may be useful for background on filtering peptide identifications. Yang's article describing the DBParser software for protein assembly explains the problems caused by protein homology (PubMed ID 15473689).

Materials and Methods

IDPicker has been implemented in C++ and C#. It is divided into three tools:

1. validateSqtToXml: Read, filter, and report peptide identifications from SQT files to XML files.
2. assemblePeptidesXml: Apply a hierarchy to collate multiple XMLs to a single XML.
3. idpickerWrapper: Filter protein identifications, apply parsimony, and report to HTML.

These three tools each implement a command-line interface. A separate program, "IDPicker Picker," was created by the Mass Spectrometry Research Center Bioinformatics group to ease the use of these programs for bench experimentalists.

IDPicker is most often used in the context of forward-and-reversed database searching, though it also supports the target and decoy protein strategy. In this strategy, each protein sequence is found twice in the sequence database, once in the normal orientation and again reversed N-to-C terminus. IDPicker

[1] This copyright to this document is owned by Vanderbilt University, 2007.

can then use matches to reversed sequences to estimate the distribution of random match scores. Peptide identifications from a SQT file are separated by charge and sorted by their absolute scores (XCorr in the case of Sequest, random match probability in MyriMatch, or hyperscore in X!Tandem). Starting from the highest-scoring match, the software assesses the peptide identification false discovery rate (FDR) that would result if all matches scoring more highly than this one were accepted. The software finds the lowest score threshold that would achieve a target FDR (usually 5%), and it accepts peptide identifications above that threshold. The peptide identifications are then sorted by their relative scores (DeltCN in the case of Sequest and MyriMatch, or reciprocal of expectation value in X!Tandem). The score thresholds are then determined for this ordering. The absolute FDR stored for each peptide reflects the resulting peptide FDR if this peptide's score were the lowest accepted, and the relative FDR reflects the result if the sort order were based on this secondary score. Mapping original scores to FDRs is managed by the validateSqtToXml tool. This is typically the slowest step in IDPicker, but it must be performed only once on each SQT file, without regard for how this file will be grouped later.

The ability to apply hierarchy in experimental reporting is a special feature of IDPicker. All IDPicker reports start with the "root" level of hierarchy, designated by the "/" label. This level encompasses all peptide identifications in the experiment. If a user is evaluating twenty SQT files divided equally between two MudPit experiments, he or she can specify this organization to IDPicker so that it groups the files appropriately. The software supports arbitrary numbers of levels in the applied hierarchy. In the above example, the user might create a hierarchy that included these labels:

- /MudPit1
- /MudPit2

In response to these labels, IDPicker would report results for each MudPit separately and also for the aggregate analysis. If the user had run four MudPits, with two replicates for each of two samples, the user might instead use labels like these:

- /Sample1/Replicate1
- /Sample1/Replicate2
- /Sample2/Replicate1
- /Sample2/Replicate2

In this hierarchy, the software would report results for each replicate of each sample, each sample in aggregate, and all files in aggregate. The user could also look for variations between the early run of the samples and the later runs by using this hierarchy instead:

- /Replicate1/Sample1
- /Replicate1/Sample2
- /Replicate2/Sample1
- /Replicate2/Sample2

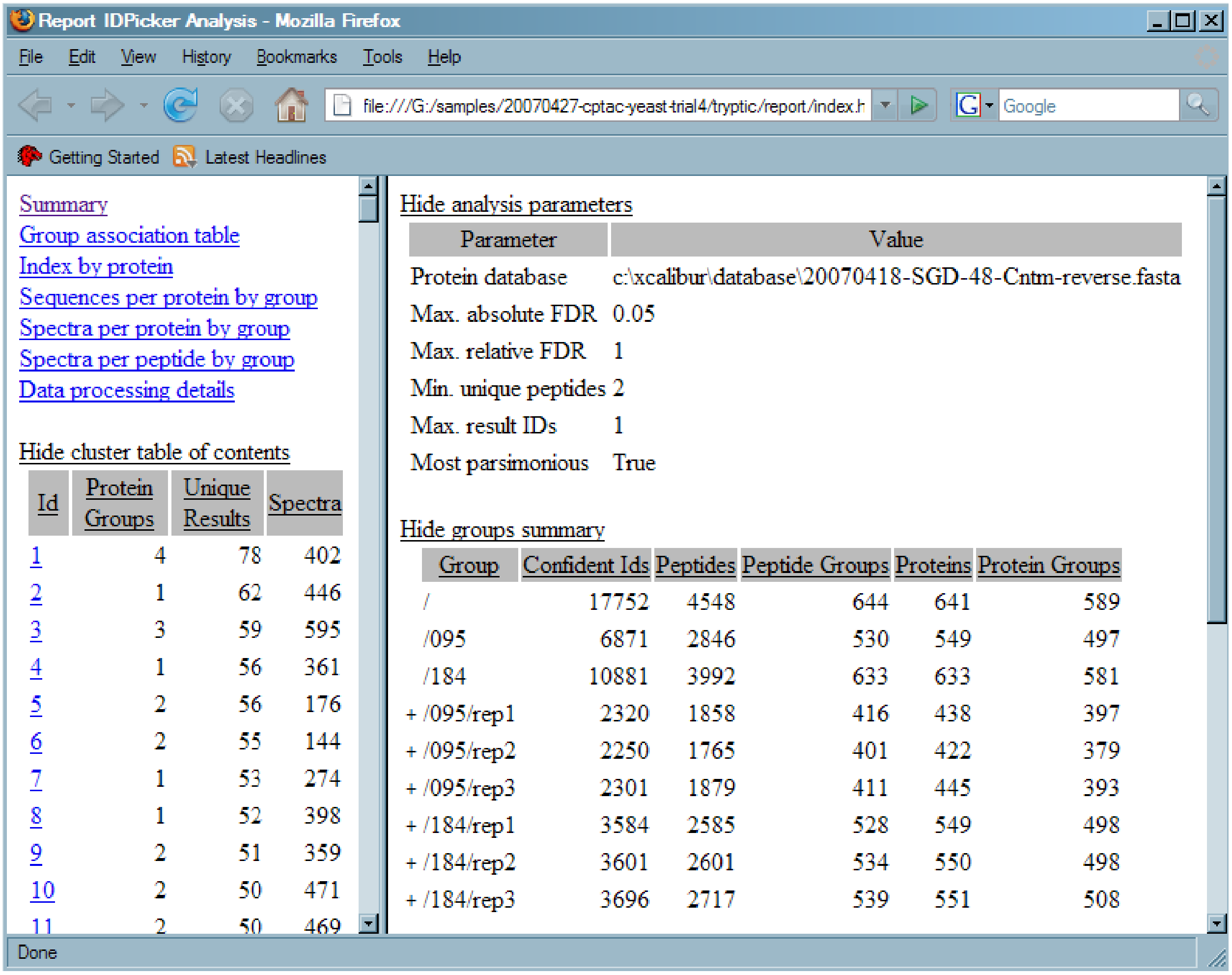
Proteins are filtered in IDPicker on two bases: parsimony and peptide count. Parsimony is an effort to produce a minimum list of proteins that accounts for all peptides accepted as legitimate identifications. In brief, IDPicker attempts to “spend” no more proteins than necessary to account for the observed peptides. The software begins with the maximal list of proteins, and it groups together the proteins pointing to the same sets of peptides. Likewise, it groups together peptides that match to the same sets of proteins. It then uses a greedy iterative algorithm to add protein groups to the final list that explain the most peptide groups that are not yet accounted for by the final list. A protein group that shares at least one peptide with another protein group will be clustered together with the other. These protein clusters are the central unit of organization used in IDPicker reporting. Using parsimony helps to limit the over-reporting of protein identifications in organisms with complex genomes.

Filtering proteins by peptide count is an effective way to limit false discovery rates at the protein level. Most typically, IDPicker is configured to remove proteins for which only one peptide has been observed. The software works upon the number of different peptide sequences observed for a protein rather than the number of spectra for that protein. When millions of spectra have been identified against a database, however, it may be necessary to require three or more peptides to keep false discovery rates under control.

Results and Discussion

The home screen of IDPicker can be reached by opening “index.html” in a web browser. The screen is divided into left and right panes. The left side shows a list of available overall reports at the top followed by a list of cluster reports. The right side shows the parameters used to create this report, some statistics for each part of the experiment hierarchy, and a histogram to show the number of clusters containing different numbers of peptides.

In this case, six different RPLC runs are being compared. Three were 95 minute separations, and three were 184 minute separations. Statistics are shown for each run individually, for all three runs of each separation type, and for all runs in aggregate. The number of different peptide sequences observed in all three 184 minute separations (3992) is smaller than the sum of the three replicate runs (2585, 2601, and 2717); many peptides observed in individual replicates are shared with other replicates. The number of peptide identifications from each replicate (3584, 3601, 3696), however, can be added together to compute the number of peptide identifications for all 184 minute separations (10881). IDPicker groups peptides into “Peptide Groups” when they match to exactly the same proteins. “Proteins” are identified sequence database entries, while “Protein Groups” are sets of proteins that are indiscernible on the basis of the observed peptides. This table is intended to give a quick overview of identification success; a replicate that has substantially different numbers of identifications can be spotted quickly by review of this table.


Report IDPicker Analysis - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
file:///G:/samples/20070427-cptac-yeast-trial4/tryptic/report/index.h
Google
Getting Started
Latest Headlines
Done

Summary
Group association table
Index by protein
Sequences per protein by group
Spectra per protein by group
Spectra per peptide by group
Data processing details

Hide cluster table of contents

Id	Protein Groups	Unique Results	Spectra
1	4	78	402
2	1	62	446
3	3	59	595
4	1	56	361
5	2	56	176
6	2	55	144
7	1	53	274
8	1	52	398
9	2	51	359
10	2	50	471
11	2	50	469

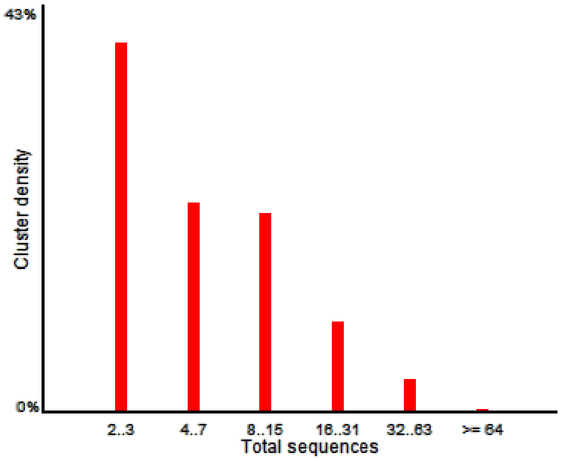
Hide analysis parameters

Parameter	Value
Protein database	c:\xcalibur\database\20070418-SGD-48-Cntm-reverse.fasta
Max. absolute FDR	0.05
Max. relative FDR	1
Min. unique peptides	2
Max. result IDs	1
Most parsimonious	True

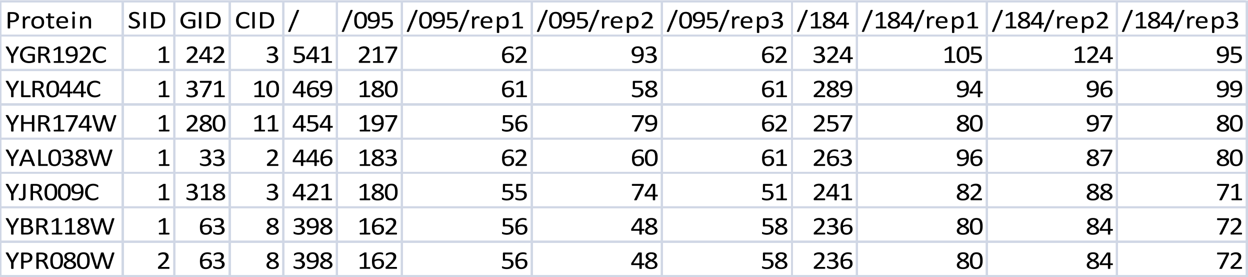
Hide groups summary

Group	Confident Ids	Peptides	Peptide Groups	Proteins	Protein Groups
/	17752	4548	644	641	589
/095	6871	2846	530	549	497
/184	10881	3992	633	633	581
+ /095/rep1	2320	1858	416	438	397
+ /095/rep2	2250	1765	401	422	379
+ /095/rep3	2301	1879	411	445	393
+ /184/rep1	3584	2585	528	549	498
+ /184/rep2	3601	2601	534	550	498
+ /184/rep3	3696	2717	539	551	508

The histogram beneath the hierarchical summary table (right) is designed to show how many peptides are found in the set of protein clusters. This image is an SVG graphic; users of FireFox will be able to see it by default, but Internet Explorer requires a plug-in from Adobe to see the picture (http://www.adobe.com/svg/). Because proteins with only two peptides are far more common than those with high sequence coverage, the histogram uses bins that double in size as they increase. If "one hit wonders" are included in the report, they will dominate the histogram.



IDPicker produces reports that enable the comprehensive tracking of proteins throughout the experimental hierarchy. From the home screen, users can click the "Spectra per protein by group" link at the upper left to produce a report like the following:



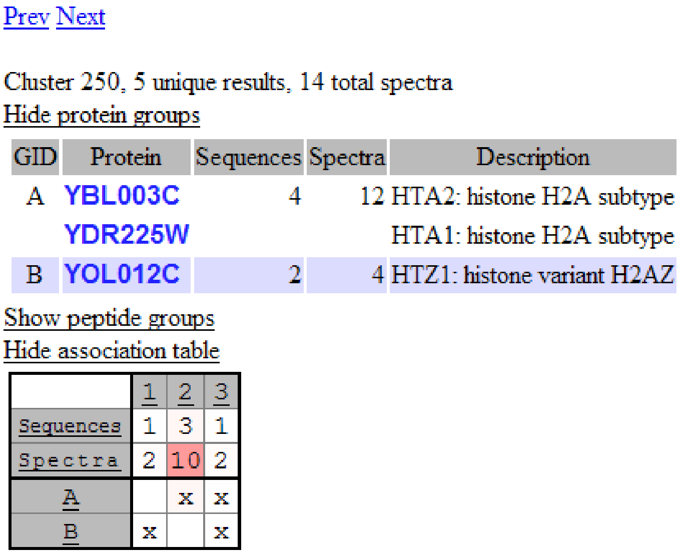
Protein	SID	GID	CID	/	/095	/095/rep1	/095/rep2	/095/rep3	/184	/184/rep1	/184/rep2	/184/rep3
YGR192C	1	242	3	541	217	62	93	62	324	105	124	95
YLR044C	1	371	10	469	180	61	58	61	289	94	96	99
YHR174W	1	280	11	454	197	56	79	62	257	80	97	80
YAL038W	1	33	2	446	183	62	60	61	263	96	87	80
YJR009C	1	318	3	421	180	55	74	51	241	82	88	71
YBR118W	1	63	8	398	162	56	48	58	236	80	84	72
YPR080W	2	63	8	398	162	56	48	58	236	80	84	72

This table, cropped to the first few proteins, shows the number of spectra matched to each of these proteins. For example, YGR192C was the yeast protein identified by the most spectra in these six experiments, with 217 of its 541 peptide identifications stemming from the 95-minute separations.

The CID, GID, and SID columns describe the relationships among proteins. CID is the “cluster identifier.” IDPicker clusters together proteins that share observed peptides. If protein A shares a peptide with protein B, and protein B shares a different peptide with protein C, all three of these proteins will be grouped together in a single cluster. In the above table, YGR192C (TDH3) and YJR009C (TDH2) are grouped together in cluster number three. Proteins that are clustered together are often related functionally, and these proteins are no exception. This cluster includes the glyceraldehyde-3-phosphate dehydrogenase enzymes of the yeast. For proteins to have the same cluster identifier is to know that at least some of their observed peptides are shared among multiple proteins.

One might worry that because YGR192C and YJR009C share peptides, one of the pair might adequately explain the observed peptides. These two proteins, however, have different GID numbers. “Group identifiers” are used to group proteins that are indiscernible on the basis of observed peptides. Because observed peptides are specific to both of the two proteins, they receive different GID numbers. This is not true, however, of YBR118W and YPR080W. These proteins (TEF2 and TEF1) are identified by 398 spectra, but none of these peptide identifications are specific to one of these two sequences. The two proteins receive the same GID as a result. IDPicker gives them different “sequence identifiers” (SIDs) to differentiate the two indiscernible proteins. If users want to have each indiscernible protein group represented only once in the spreadsheet, they can sort by the SID column to push these bonus identifiers to the end of the sheet.

The cluster report is the main means by which IDPicker describes protein / peptide relationships. These reports are available by clicking the numbered links in the left pane of the home screen (or by a search for a particular description in the “Index by protein”). Clusters are sorted by the numbers of different peptide sequences they contain and then by the numbers of spectra they represent, and users can navigate to the previous or next cluster by following the links at the top of the cluster report. In this case, the cluster contains three histone H2A proteins. The first two identifiers, YBL003C and YDR225W, cannot be discerned on the basis of observed peptides. To reflect this, IDPicker shows them both as part of group A with a white background. Group B is differentiated by the shaded background and by the sequence and spectral counts reported by the identifier.



Prev Next

Cluster 250, 5 unique results, 14 total spectra

Hide protein groups

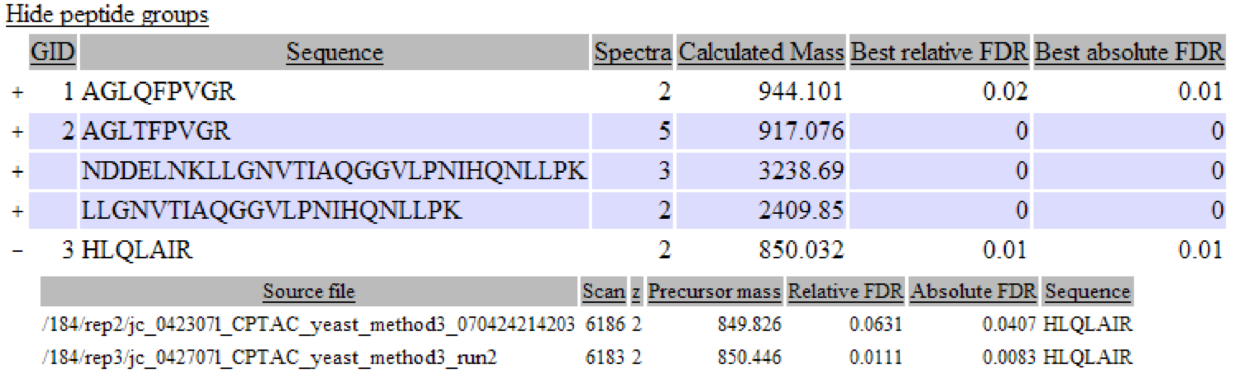
GID	Protein	Sequences	Spectra	Description
A	YBL003C	4	12	HTA2: histone H2A subtype
	YDR225W			HTA1: histone H2A subtype
B	YOL012C	2	4	HTZ1: histone variant H2AZ

Show peptide groups

Hide association table

	1	2	3
Sequences	1	3	1
Spectra	2	10	2
A		x	x
B	x		x

The proteins in group A and B are related by shared peptides. The peptide association table below the protein table shows this relationship. If only one protein is found in a cluster, the association table is omitted. In this example peptide group three (represented by the final column of the table) contains one sequence that was matched to two different spectra. This peptide group matches to proteins in both groups A and B. There is, however, evidence specific to each of these protein groups. The peptide sequence in group 1 (second column of the association table) was matched to two spectra and is specific to the variant histone protein of group B. The three peptide sequences of group 2 (matched to ten spectra) can be matched only to the proteins of group A. If a user clicks on the "B" in the final row of the table, IDPicker will highlight the peptide columns that match this protein. As larger numbers of proteins are clustered together, these tables can grow considerably, but the principles shown in this example are the same for a larger table.



Hide peptide groups

	GID	Sequence	Spectra	Calculated Mass	Best relative FDR	Best absolute FDR
+	1	AGLQFPVGR	2	944.101	0.02	0.01
+	2	AGLTFPVGR	5	917.076	0	0
+		NDDELNKLLGNVTIAQGGVLPNIHQNLLPK	3	3238.69	0	0
+		LLGNVTIAQGGVLPNIHQNLLPK	2	2409.85	0	0
-	3	HLQLAIR	2	850.032	0.01	0.01

Source file	Scan	z	Precursor mass	Relative FDR	Absolute FDR	Sequence
/184/rep2/jc_0423071_CPTAC_yeast_method3_070424214203	6186	2	849.826	0.0631	0.0407	HLQLAIR
/184/rep3/jc_0427071_CPTAC_yeast_method3_run2	6183	2	850.446	0.0111	0.0083	HLQLAIR

Users can retrieve the full list of peptide identifications for each protein cluster by clicking the "Show peptide groups" link. The peptides for the histone H2A cluster look like the above. The sequence for each peptide is shown along with the number of matched spectra, the computed peptide mass, and the false discovery rates associated with that peptide's best absolute and relative scores. In this example, the "+" symbol by the final peptide was clicked to show the specific spectra identified to this peptide.

Conclusion

IDPicker affords considerable flexibility to users, enabling the summarization of complex data sets. Because different proteomic applications require different report styles, IDPicker has been designed to accommodate multiple reporting and leave room for added connectivity to other tools. Likewise, compatibility with multiple search engines has been a priority during development, and the software works with any identifier for which results can be translated to SQT format. IDPicker reports the shared peptide relationships among identified proteins to enable a more nuanced interpretation of protein identifications.

Acknowledgment

Matthew C. Chambers implemented the code of IDPicker. Bing Zhang designed the bipartite graph representation upon which the parsimony features of IDPicker were based. This work was funded by R01 CA126218 and U24 CA126479 from the National Cancer Institute.