Mascot Distiller 2.4: Quantitation of Very Large Datasets

MASCOT

{MATRIX \ {SCIENCE}



When we added support for quantitation to Mascot Distiller, we failed to anticipate how many people would want to process large collections of raw files as a single experiment. The original software architecture was designed to read large chunks of data into memory to make processing fast. Because it was a 32-bit application, which cannot use more than 2 GB of address space, this meant it would crash if you had a project where the total size of the raw files was more than 5 GB or so.

Making a 64-bit version of Distiller removes this limit but, for many people, this would mean moving to a new version of Windows and possibly new hardware. Also, unless you have a huge amount of RAM, a very large project would run out of RAM even though it hadn't run out of address space. The real fix also requires the processing to be serialised so as to keep as much of the data as possible on disk



Distiller 2.4 is available in both 32-bit and 64-bit versions. This means that you aren't forced to move to a new PC immediately, although 64-bit is the future and is required for very large projects

When you create a multi-file project, the default is to process the files independently. If you want the old behaviour, to process all the data as if it was in one huge raw file, this is still available as an option. By default, peak picking, database searching, and quantitation are completely independent. The search results from the individual files are merged into a single list of proteins, because we need this to organise the quantitation results. This is the point at which you might run into problems with the 32-bit version if the data set is extremely large.

When browsing search or quantitation results, only one raw file is open at a time, so there is a slight delay when you move between files.



I'd like to show how it works by walking through some screen shots. The dataset is a set of public domain Orbitrap SILAC files. There is a link on the maxquant.org site to download the files from Proteome Commons. We used a reasonably high spec PC. First, lets look at processing everything in Distiller workstation



As in 2.3, we choose to create a new multifile project

Wascot Distiller	
File View Processing Analysis Tools Help	
Image: Dec Processing Analysis 1000 mpp ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	
	T. C.
Ready	Info

There is a change here. The dialog allows us to choose raw files or existing projects or a mixture. We'll look at using existing projects later.

Ascot Distiller	_ 🗆 X
File View Processing Analysis Tools Help	
☆ 金白白 ↓ ↓	
Acquisition V 4 X	
Select XCalibur data set	
Name To Date modified Type	
Recent Places 20070522_NH_orbi2_HelaEpo_02.RAW 12/01/2008 19:52 RAW File	
20070522_NH_0rbi2_HelaEpo_03.RAW 12/01/2008 19:52 RAW File	
20070522_IMI_orbit2_HeleEpo_04.RAW 12/01/2008 19:49 RAW File	
20070522_HI_0rbt2_Helabo_0.6.RAW 12/01/2008 19:45 RAW File	
20070522_NH_0rbi2_HelaEpo_07.RAW 12/01/2008 19:43 RAW File	
Libraries 20070522 JNH_Orbi2_HeldEpp_08.RAW 12/01/2008 19:41 RAW File	
20070522_HI_OrbiZ_Helebo_10.RAW 12/01/2008 19:37 RAW File	
Computer 2007/0522_NH_0rbi2_HelaEpo_11.RAW 12/01/2008 19:36 RAW File	
20070522_JNI_obt2_HeleEpo_12.RAW 12/01/2008 19:11 RAW File	
Network	
File name: "20070611 NH Orbi2 HeleEpo 24 BAW" "21 Deens	
Files of long VC should and Cancel	
Thesion type: [Accallular [, 1849]	
Help	
Processing Option	
Field Value	
Dearty	Info

For now, we browse to the raw files and select all $72\,$

🙀 Mascot Distiller		
File View Processing Analysis Tools	Help	
📸 🚖 🖆 🔁 🐰 🗅 🗂 🖄 🚯 💕	じる古れいの目 兵路後 妖 局局局局 山北 後米や路沙 君毛	
Acquisition 👻 🕂 🗙		
	New Multi File project	
	Sample name # Status Filename Add File(s)	
	64 1 New 20070611_NH_Orbi2_HelaEpo_	
	65 1 New 20070611_NH_Orbi2_HelaEpo_	
	67 1 New 20070611_NH_Orbi2_HelaEpo Add Project	
	68 1 New 20070611_NH_Orbi2_HelaEpo_	
	69 1 New 20070611_NH_Orbi2_HelaEpo_	
	70 1 New 20070611_NH_Orbi2_HelaEpo_	
	71 1 New 20070611_NH_Orbi2_HelaEpo	
	Process Independently (Not compatible with Replicate)	
	Cancel	
	Distiller Project File	
	C:\data\MaxQuant\NH_Orbi2_HelaEpo.rov	
	Processing Options	
	Orbitrap high res MS2 Open	
	LCQ without zoom scans	
	LTQ FT	
	No peak picking, take M52 centroids from raw file	
	Hybrid, peak pick MS1 and MS2 up to 7+ Hybrid, peak pick MS1 up to 4+ and MS2 up to 2+	
	Hybrid, peak pick MS2 only Hybrid, peak pick MS1 and take MS2 control is from raw file	
	Default options	
Ready		Info

The selected files are listed. You can add further files or remove files if you change your mind. This is where you choose a name for the project and a chance to select the processing options for peak picking. When everything looks OK, choose Open



You'll notice that one file, the active one, shows on the acquisition tree as an Xcalibur icon with a TIC. The others, that are not currently in memory, show as hour glass icons. If you wanted to browse the raw data, and click on one of the files with an hourglass, there would be a few seconds delay while the first file was closed and the selected file was opened. In this walk through, we'll go directly to peak picking and database searching



You specify the peak picking via the processing options file. For routine work, you will simply choose one that has been optimised for the type of data. The search conditions can be saved as part of the project preferences or entered into a search form at the point of submitting a search. These are the conditions we will use for this dataset



Peak picking and database searching can be performed for all files in the project by choosing 'Process and Search'



This is where you'll see a substantial improvement in speed compared with Distiller 2.3. The code is threaded so that all the processor resources can be used. Initially, the progress dialog looks like this

Mascot Distiller - NH_Orbi2_HelaEp	*
File Edit View Processing Analys	is Tools Windows Help
🖆 🚖 🖆 🛱 🗶 🛍 🗂 🛬 🕄	ジ ២ 0 凸 加 μ 刀 圖 文 苔 茨 爻 局 ゐ क़ 昂 山 地 氷 ※ ∾ 路 ≫ 宮 등
Acquisition + # ×	NH Orbi2 HelaEpo* - 20070522 NH Orbi2 HelaEpo 01.RAW
🖃 🥘 NH_Orbi2_HelaEpo 🔺	
20070522_NH_0rbi:	Process and Search
₩	Processing complete 23 of 72
20070522_NH_0rb1:	
20070522 NH Orbi:	Search complete 5 of 72
20070522_NH_0rbi:	
-X 20070522_NH_0rbi:	20070522_NH_Orbi2_HelaEpo_23 (Searching for peaks)
-X 20070522_NH_0rbi:	
20070522_NH_0rbi:	20070528_NH_Orbi2_HelaEpo_01 (Searching for precursor peaks)
20070522_NH_0rb1	
20070522_NH_01D1.	2007/05/28 NH Orbi2 HeleFop 02 (Searching for precursor peaks)
20070522_NH_Orbi:	
-X 20070522_NH_0rbi:	20070520 NUL O-Hi2 Hale Eng. 02 (Example a fer average and a)
- 20070522_NH_0rbi:	zovozo_jwi_oruz_ielacho_o (bearding for precursor peaks)
- 20070522_NH_0rbi:	
20070522_NH_0rb1:	20070528_NH_Crbi2_HelaEpo_04 (Searching for precursor peaks)
20070522_NH_0rb1	
20070522 NH Orbi:	20070522_NH_Orbi2_HelaEpo_05 (Searching)
20070522_NH_0rbi:	
-X 20070522_NH_Orbi:	20070528_NH_Orbi2_HelaEpo_05 (Searching for precursor peaks)
-X 20070522_NH_0rbi:	
- 20070522_NH_Orbi:	20070528_NH_Orbi2_HelaEpo_06 (Collecting precursor information)
20070522_NH_0fb1	
20070528 NH Orbi:	20070528 NH Orbi2 HelaEpo 07 (Open datasource)
20070528_NH_Orbi:	
-X 20070528_NH_0rbi:	
-X 20070528_NH_0rbi:	Details
20070528_NH_0rbi	
20070528_NH_0rb1	- I
Acquisition Precursors	500 5000 6500 7000 7500 3000 3500 4000 4500 5000 5500 6000 6500 7000 7500 Sec
Ready	

A short while later, we are about 1/3 the way through peak picking and 5 searches have been completed



In this example, peak picking and searching all 72 files took 7 hours 45 minutes to complete. The search results have been merged into a single, minimal list of proteins. You'll notice that the proteins tab now uses the new family grouping, introduced in Mascot 2.3. This is particularly useful for quantitation because it ensures that proteins related by shared peptide matches are displayed together, making it easy to spot whether a particular peptide match should belong to one isoform rather than another. Family grouping is an option; you can choose the Select Summary-style list if you prefer it.

The next step is quantitation. You can process some or all of the proteins. We'll choose all.



The first step is to collect together peptide matches from all the search results that correspond to the same sequence.

Kascot Distiller - NH_Orbi2_HelaEpo*	_ 🗆 X
File Edit View Processing Analysis Tools Windows Help	
🗃 👌 🖆 🖥 🖞 🖞 🗳 🕑 野│ ២ 🖯 📇 🗽 🎵 🥅 瓜 🏯 🎊 💢 📑 வுளை வுளை 🖬 🛄 🛝 🕸 🎭 🤮 🚝 🜉	
Proteins • 4 X	_ 🗆 ×
- WH_Orbi2_HelaEpo	
- K Format Options	**//
Search header	
Quantitation Progress (24 threads)	
Data Sources :	
Solution match:2+ AFAFSMYOIX	
2.2 gill4657944	
2.3 gi 119569035	
3.1 gi131542947 1 2 :Quantitation match:2+,IVIDSNNNPER	
🖶 🎆 3.2 gi 222476522	11
a 🙀 4.1 gil4501887 H 3:Quantitation match:2+,SGDAAIVDMVPGK	
₽ 🙀 4.2 gi 168177284	
⊕ % 4.4 gi1221043300	
4.5 gill8505325	
4.5 g11644211/0 J 5:Quantitation match:3+,AYLPVNESF6F1ADLR	
Grossian for the second s	
+ Statistics match: 3+.VSHVSTGGGASIELLEGK	
5.1 gill2667788 1	
5.2 gill09734611	
5.3 gi 116284394 8:Quantitation match:3+,AEDGSVIDYELIDQDAR	
🕂 🎆 5.4 gi 119610406	
🖶 🎆 6.1 gil4503571 🕅 Details Cancel	
6.2 gi 254763441	
6.3 gil221040676	
	mm 1
Peak Lists Proteins Searches	n 7500 See
Acquisition Precursors	0,300 380
Ready	11.

Once quantitation gets going, it also uses multiple threads. This system has dual 6-core processors, and each core is hyperthreaded, so 24 threads are used. Even so, it takes some 22 hours to quantitate all 4376 proteins

Mascot Distiller - NH_Orbi2_Hel	aEpo*									
File Edit View Processing Analysis Tools Windows Help										
22 22 24 12 12 12 12 12 12 12 12 12 12 12 12 12										
Proteins • 4 ×	Proteins • I X Quantitation Table • I X					→ ‡ ×				
🕞 🥝 NH_Orbi2_HelaEpo 🔺		Accession	Score	Mass	L/H	SD(geo)	#		Description	<u>م</u>
	1.1	gi 116063573	186258	282581	1.2893	1.3075	3321	filamin-A isoform 1 [Homo	sapiens]	
Search header	1.2	gi 256222411	111954	283626	1.2533	1.5700	2209	filamin-B isoform 1 (Homo	sapiens]	
init and a first the first the first the first the first term of term o	1.3	gi 34365016	88529	259170	1.2692	1.5702	1889	hypothetical protein (Homo	o sapiens]	
±∰ 1.2 gi 256222411	1.4	gi 116805322	19414	293407	1.2380	1.5852	306	filamin-C isoform a [Homo	sapiens]	
±∰ 1.3 gi 34365016 l	2.1	gi 35505	149397	58412	1.2595	1.4620	1677	pyruvate kinase [Homo sa	ipiens]	
1.4 gi 116805322	2.2	gi[114657944	113670	62125	1.2663	1.4932	1545	PREDICTED: pyruvate kina	ise 3 isotorm 2 (Pan ti	rogioaytesj
1 2.1 gi 35505 Mas:	2.3	gij119569035	146503	59000	1.1305	1.20/9	1249	RCG401131, ISUIUMI CRA	_p [Huniu sapieris]	conienel
	3.2	gij31342847	587	19880	1 3311	21 9074	47	60 kDa chaneronin (Homo	saniens)	
1.3 gi 119569035	1	- gazzati obzz	001	10000	1.0011	21.0074			ouplotioj	
3.1 gi 31542947]										
3.2 gi 222476522	MINH_	Orbi2_HelaEpo* - :	20070522_NH_	Orbi2_HelaEpo	_01.RAW					
4.1 ct 14501887 M										A
4 2 mi 1168177284										10/L
± 4.3 mi 134133226										
4 A gi 221043300										
4.4 g1 221040000										
4.5 g1 11000325										
4.0 gr 024211/0 1										
4.7 g1 62420949 1										
4.8 g1 88953571 1										6 11
4.9 g1[62420916]										
4.10 g1 5031571 J										
4.11 gi 223597 M										
⊕.∰ 5.1 gi 12667788 1										
⊕ 🐝 5.2 gi 109734611										
🕂 👯 5.3 gi 116284394										
⊕∰ 5.4 gi 119610406										- B 10
🖶 🎆 6.1 gi 4503571 M										
⊕∰ 6.2 gi 254763441										- (1 11
±∰ 6.3 gi 221040676										
±∰ 6.4 gi 119612018										
🕂 🍇 7.1 gi 7106439 M										- 1011
+ ¥ 7.2 gi 23958133 1▼		ļ			Juli	huhurm	num	human		
PeakLists Proteins Searches										
Acquisition Precursors		500 100	U 1500 20	UU 2500 3	UUU 3500	4000 4500	5000	5500 6000 650	U 7000 7500	Sec
Ready										

When complete, it looks much the same as before apart from the family grouping. There is a change in the way the quantitation table is displayed. You can now choose between having peptide rows indented in the main table or displaying peptides as a separate, linked table.



The linked table, as shown here, usually works better for very large tables. You'll notice that the peptide matches adjacent to the selected one are in grey. This is a visual clue that they come from a different raw file, and if you click on one of them, there will be a short delay while the files swap over. Peptide 1418 is from 01.raw



While 1417 is from 09.raw

After you save the project, its reasonably fast to reopen. This particular one takes just under 3 minutes to open from disk. Remember that, if Distiller is not registered, it operates in viewer mode, so this is a powerful way of sharing search and quantitation results with colleagues



The other workflow is to use Mascot Daemon to batch process the individual files. This requires Distiller to include the Daemon Toolbox option, so that Distiller can be called by Daemon. Daemon automates all of the processing steps, from peak picking to quantitation, and saves each project to disk. You can then open the set of projects in Distiller Workstation to create a multi-file project where all the data processing for the individual raw files has already been completed

Status	Event Log	Task Editor	Parameter Editor	
· · · ·		_	·	
- Owner	Task			
	Mascot Daemon: Data import	filter options		🗵 Run
, <u> </u>	Icg_dta.exe Micromass.spl	Sciex Analyst Data Explorer	Mascot Distiller TS2Mascot	
Parameter set			1 1	
C:\Program Files (x86)\M	Mascot Distiller Processing C)ptions		Options
Data Galla	C:\MaxQuant\maxquant.o	pt	Edit Save As	
D'ata me iist	Data File Format	Text Data File (MS o	n(y) Intensity values	
Drag and drop data files in Add	1 Thermo Xcalibur	C Centroided pe	sak list 💿 Area	
		Profile / conti	nuum O S/N	
ant\MaxQuant\20070611 ant\MaxQuant\20070611	Multi-Sample Files	Quantitate F	Protein Hits	
ant\MaxQuant\20070611	 Merge all samples into s Separate search for earch 	single search (All C Bange	rom to	
ant\MaxQuant\20070611	Care David (m. K. and Ga) (Trange	Ded Mit Ferret	
ant\MaxQuant\20070611	Start Mange (Induk-scan nies	J		rocesses
ant\MaxQuant\20070611		I Onics Minutes		
ant\MaxQuant\20070611 ant\MaxQuant\20070611	Output PMF Masses as	Output MS/MS Fragm	ents as Distiller Project File	
•	Om/z ⊙MH+ O	Mr © m/z O MH+	C Mr 🔽 Save	
Add				
Merge MS/MS files int	to single search	Pass data to	None	<u>~</u>

Daemon is great for routine work. You don't have to remember any settings. Just clone a previous task and change the list of files. Make sure the box is checked to save the project and choose to quantitate all hits, because there is no guarantee that hit 10 in an individual file will be hit 10 in the merged search results

Mascot Distiller	_ _ ×
●●●「★「」」●●●●●「「●●」」●●●」(●●●」(●●●●」(●●●●」	
Acquisition • 4 ×	
New Multi File project	
Samula nama # Status A add Ele(-)	
1 New C~~data-MaxQuant~MaxQuant~	
2 1 New C~~data~MaxQuant~MaxQuant~	
3 1 New C-wodata-MaxQuant-MaxQ	
4 1 New C-watar-MaxQuant-MaxQuant-	
6 1 New C~~data~MaxQuant~MaxQuant~	
7 1 New C~~data~MaxQuant~	
8 1 New C-watata-MaxQuant-MaxQuant-	
Process Independently (Not competible with Replicate)	
Cancel	
Distiller Project File	
C:(MaxQuant)Maxquant 125.rov	
Processing Options	
Open	
Ready	//.
MASCOT : Mascot Distiller 2.4 © 2011 Matrix Science	
	ENCEJ

Once the Daemon tasks are complete, we can select the project files for the multi-file project



You still choose Process and Search. This checks that the peak picking and search conditions are identical across all the projects, then extracts the search results. If any of the projects used different peak picking or search settings, they would be re-processed and/or re-searched during this procedure. As would any raw files that had been included.

🗱 Mascot Distiller - Maxquant 125.rov*			
File Edit View Processing Analysis Tools Windows			
	₩♡╨☆╨₩∞₽■	四目書が発きる取る「取用」	
Dearches • • • X	/* - C:\data\MaxQuant\MaxQuant\20070522_	NH_Orbi2_HelaEpo_01.RAW	
Master Search 1			
Submitted from			
Submitted from			
Submitted from			
UL Submitted from	Peptide Summary		
Submitted from Creates	ummary		
Jubnitted from			
Grouping Grouping	peptides		
Submitted from			
Submitted from	die Canaal		
Submitted from			
Submitted from			
Submitted from			11
Submitted from			
- II Submitted from			
Submitted from			
Submitted from			
Submitted from			6
II Submitted from			
Submitted from		t t a statut	V
	_rlunhunhunh	miles when her bell have men	mannah 1
	1000 1500 2000 2500 3000 3500	4000 4500 5000 5500 6000 65	00 7000 7500 Sec
Acquisition Precursors beartnes			7
Incody			1.0
MASCOT : Mascot D	istiller 2.4	© 2011 Matrix Science	(MATRIX) SCIENCES

Creating the combined peptide summary is the most time consuming step



Quantitation is just a case of extracting the existing data from the individual project file



The final result is exactly the same as if we started from raw files, but the time taken for this particular example was under 2 hours. You can easily remove projects or add new ones, as long as they were processed using identical settings. This is ideal for experimenting with replicates. You can look at the results for the individual replicates and the combined results without having to start from scratch every time.



Besides fixing the memory and speed problems for multi-file projects, there are a number of other new features. The 2.4 release will bring Distiller back into line with a couple of features that were new in Mascot Server 2.3: protein family grouping and re-scoring using Percolator. (Although, you cannot use Percolator for multi-file projects.) By popular request, you can now export de novo solutions as an XML file. Some of the data format libraries have been updated, including the Bruker libraries, so we can finally open maXis data.

The ability to open mzML files is particularly important for AB Sciex TOF-TOF data. Previously, there was no easy way to work with this data in Distiller because it is stored in tables in an Oracle database.



Recently, AB Sciex developed a utility that can export a spot set as an XML file, with conventional parent child relationships between the MS and MS/MS scans. This is still in beta test, so I can't provide a download link at this time.



Here is a small mzML file of 4800 TOF-TOF data, courtesy of Ida Chiara Guerrera and François Guillonneau, Université Paris Descartes. It looks just like conventional LC-MS/MS data when opened in Distiller



Which opens up the possibility of quantitation using precursor protocol methods such as the SILAC experiment shown here

Let me anticipate the first question: When can I get it? Currently, we are in beta test. All being well, beginning of July