

# DTASelect FAQ

David L. Tabb

March 18, 2002

## Contents

<b>1 DTASelect</b>	<b>2</b>
1.1 Why use DTASelect at all? . . . . .	2
1.2 Why are we throwing out all these matches? . . . . .	2
1.3 Error messages . . . . .	2
1.3.1 X is garbled! . . . . .	2
1.3.2 X is missing information! . . . . .	2
1.3.3 X did not list any matches! . . . . .	2
1.3.4 Sequence for X does not match Y sequence . . . . .	2
1.3.5 X has XCorr of 0.0! . . . . .	2
1.3.6 X reports no sequence! . . . . .	2
1.3.7 Compressed file not found. . . . .	3
1.4 Graphical User Interface . . . . .	3
1.4.1 The GUI sequence coverage viewer doesn't show uncovered sequence. . . . .	3
1.4.2 How do I print the image of a spectrum? . . . . .	3
1.5 I ran DTASelect again with different options, but the html file looks just the same. . . . .	3
1.6 My peptides are much shorter than the ones the sequence coverage program reports. . . . .	3
1.7 Some of the sequence contigs sent to the coverage viewer do not show up in the display. . . . .	3
1.8 The .out file says this peptide appears in several proteins, but it lists only one. . . . .	3
1.9 Isn't there some way to avoid retyping the same options? . . . . .	3
1.10 I moved my data to a new location, but DTASelect's links still point to the old location. . . . .	3
1.11 How should I structure my data on disk for DTASelect to combine multiple LC runs together? . . . . .	4
1.12 How can I combine multiple DTASelect results together? . . . . .	4
1.13 How do redundant and nonredundant counts differ? . . . . .	4
1.14 The reported count of proteins is lower than the number I see onscreen. Why? . . . . .	4
1.15 How can I enter my validations for proteins or peptides? . . . . .	4
1.16 How can I get the background graphic to appear on my install of DTASelect? . . . . .	4
<b>2 Contrast</b>	<b>4</b>
2.1 Where's the GUI for Contrast? . . . . .	4
2.2 I specified -GUI or -compress in my DTASelect.params, but Contrast ignored it. . . . .	5
2.3 Contrast reports a protein is present when it isn't. . . . .	5
2.4 The reported count of proteins is lower than the number I see onscreen. Why? . . . . .	5
2.5 Why is there only one set of coverage percentages for each group of proteins? . . . . .	5

# 1 DTASelect

## 1.1 Why use DTASelect at all?

I can think of five reasons:

- It's fast: it takes a few minutes to process tens of thousands of SEQUEST output files.
- It's consistent: it applies criteria uniformly and without exceptions.
- It's transferrable: if you have a set of criteria you like, you can apply them to other data and share them with others.
- It's adaptable: if you decide your first set of criteria wasn't quite right, you can try another in seconds.
- It's automated: you can spend your time doing more challenging things than poring over a pile of numbers.

## 1.2 Why are we throwing out all these matches?

SEQUEST attempts to find any database sequence that matches the spectrum. A spectrum, however, may not actually represent a sequenceable peptide. It might instead be a solvent aggregate, a peptide with an excessive or incorrectly assigned charge state, or a peptide with an unspecified modification. SEQUEST reports the best that could be done for each spectrum, but in many cases, the spectrum simply can't be interpreted without more information.

## 1.3 Error messages

### 1.3.1 X is garbled!

DTASelect was trying to read an .out file, but the output file's format appeared to be incorrect. This could mean that DTASelect expected a number but found letters instead.

### 1.3.2 X is missing information!

DTASelect was trying to read an .out file, but a line it was trying to get information from was too short.

### 1.3.3 X did not list any matches!

DTASelect ran across an .out file which didn't contain any matching peptide sequences. These .out files just report that no database sequence was at all plausible.

### 1.3.4 Sequence for X does not match Y sequence

Some database loci may have descriptions that are so long that SEQUEST interprets them as sequence. These matches are incorrect ones and will be disregarded. Another possibility is that the databases on your DTASelect machine are not the same version as the ones on your SEQUEST machine.

### 1.3.5 X has XCorr of 0.0!

If some versions of SEQUEST fail to find any matching sequence for a spectrum, they will match it to a random sequence and give it a score of XCorr. These are meaningless and are discarded.

### 1.3.6 X reports no sequence!

Some versions of SEQUEST can sometimes match a spectrum to a sequence like X..X, or effectively no sequence at all. DTASelect reports this occurrence and ignores the spectrum thereafter.

### **1.3.7 Compressed file not found.**

You've tried to view a spectrum for which the .dta file is missing and for which no DTASelect.IDX and DTASelect.SPM files are present (use the `--compress` option to create them from existing .dta's).

## **1.4 Graphical User Interface**

### **1.4.1 The GUI sequence coverage viewer doesn't show uncovered sequence.**

Since the DTASelect.txt file doesn't contain the database sequences, these cannot be filled in. The file would have to be much larger to implement this feature.

### **1.4.2 How do I print the image of a spectrum?**

Printing is not implemented in DTASelect, but Windows can take a screenshot when you hit ALT-PrintScrn. You can then use "paste" to put this image into a Word document or graphics program.

### **1.5 I ran DTASelect again with different options, but the html file looks just the same.**

Click the "Refresh" or "Reload" button for your browser. Different settings may yield the same sets of proteins, but the DTASelect.html file should still show the difference in the command line near the top of the page.

### **1.6 My peptides are much shorter than the ones the sequence coverage program reports.**

DTASelect merges together overlapping peptide sequences to shorten the sequences passed to the coverage display program. This allows very high sequence coverages to be displayed correctly.

### **1.7 Some of the sequence contigs sent to the coverage viewer do not show up in the display.**

The sequence coverage viewer shipped with SEQUEST has many flaws; some versions can only show the overlap between a single peptide and the protein sequence. In addition, the program does not always correctly align peptide sequences with the protein sequence. Use the DTASelect GUI to ensure that you are seeing correct sequence coverage.

### **1.8 The .out file says this peptide appears in several proteins, but it lists only one.**

Check to make sure that `print_duplicate_sequences` is turned on in your `sequest.params`. SEQUEST must be re-run on the sample for this to go into effect. If the other protein names are not enumerated in the .out files, DTASelect won't pick them up.

### **1.9 Isn't there some way to avoid retyping the same options?**

Yes, put the options you're typing into a file called `DTASelect.params` and copy it into the directory where you're running DTASelect (not the DTASelect installation directory).

### **1.10 I moved my data to a new location, but DTASelect's links still point to the old location.**

DTASelect.txt files can be used independently of the data which they represent; you can move a DTASelect.txt file to a new location and still get links to the old location. Yes! It's a feature, not a bug!

If you need the DTASelect.txt to make links to a new location, change the second line of the file in a text editor to the new location of spectra and SEQUEST result files.

### **1.11 How should I structure my data on disk for DTASelect to combine multiple LC runs together?**

Each LC run's SEQUEST results should be in a separate directory. These subdirectories should be named identically to the .raw files (for example, the file Yeast01.raw should be extracted and SEQUESTed in a directory called Yeast01). These subdirectories should be inside another directory (creating, for example, a directory called Yeast with subdirectories titled Yeast01, Yeast02, Yeastxx, inside it). The sequest.params for all directories should be identical, and a copy of the sequest.params should be in the parent directory (Yeast, in this example). DTASelect is run in the parent directory, creating a unified DTASelect.txt that includes SEQUEST results from each subdirectory.

### **1.12 How can I combine multiple DTASelect results together?**

Create a Contrast.params that lists all the files you want to combine. Specify "merge" in the options section, and when you run Contrast it will create a unified DTASelect.txt file in the current directory.

### **1.13 How do redundant and nonredundant counts differ?**

The nonredundant number of proteins counts proteins after grouping by identical sequence coverage; the redundant count counts each protein regardless of grouping.

The nonredundant number of peptides counts each individual spectrum only once; if the same spectrum is associated with multiple loci, it is counted multiple times for the redundant count. The spectrum counts differ in the same way. For example, if protein A incorporates peptides 1, 2, and 3 while protein B incorporates peptides 1, 4, and 5, the nonredundant number of peptides is 5 while the redundant number of peptides is 6. If peptide 1 is present in four copies, it will count for four copies nonredundantly and eight copies redundantly. If you use the "-t 0" setting, the number of spectra will be completely bogus.

### **1.14 The reported count of proteins is lower than the number I see onscreen. Why?**

See section 1.13.

### **1.15 How can I enter my validations for proteins or peptides?**

To do this easily, one would need CGIs that handled the feedback from the user's clicking on the validation letters by protein names or which would handle marking individual identifications. DTASelect is prepared to deal with four different states for any locus or identification:

- Y Yes, this is a legitimate match
- M Maybe this is legitimate
- N No, this is not a legitimate match
- U I haven't gotten to this one yet or don't care

The information is stored in the DTASelect.txt file. Each "L" (Locus) or "D" (DTA identification) line concludes with a tab character followed by the letter "U" until this is changed by an external program to a "Y," "M," or "N." The next time DTASelect is run on this DTASelect.txt file these values will be used for marked peptides.

### **1.16 How can I get the background graphic to appear on my install of DTASelect?**

DTASelect will automatically look in the "images" subdirectory of your website for a file called "marble.jpg" and use it as the background for its HTML reports. Needless to say, this file need not actually contain an image of a marble slab.

## **2 Contrast**

### **2.1 Where's the GUI for Contrast?**

It doesn't exist. GUI creation takes time that I don't have.

## **2.2 I specified –GUI or –compress in my DTASelect.params, but Contrast ignored it.**

Only DTASelect is capable of using a GUI, and so only DTASelect can compress the spectra.

## **2.3 Contrast reports a protein is present when it isn't.**

Make sure that each directory and each criteria set is given a different short name in Contrast.params. If this isn't done, the html files created will overwrite each other and confuse the results.

## **2.4 The reported count of proteins is lower than the number I see onscreen. Why?**

See section 1.13. Also, note that Contrast may show a higher nonredundant number of proteins than DTASelect because a group of proteins has been split to two groups because of presence and absence information.

## **2.5 Why is there only one set of coverage percentages for each group of proteins?**

It's a little tricky to do the lookups for each of the proteins, and since they're all closely related, it seemed simpler just to list the sequence coverage percentages for the first of them. The cumulative sequence coverages for each are in the "total" column.