Supplementary Information S2

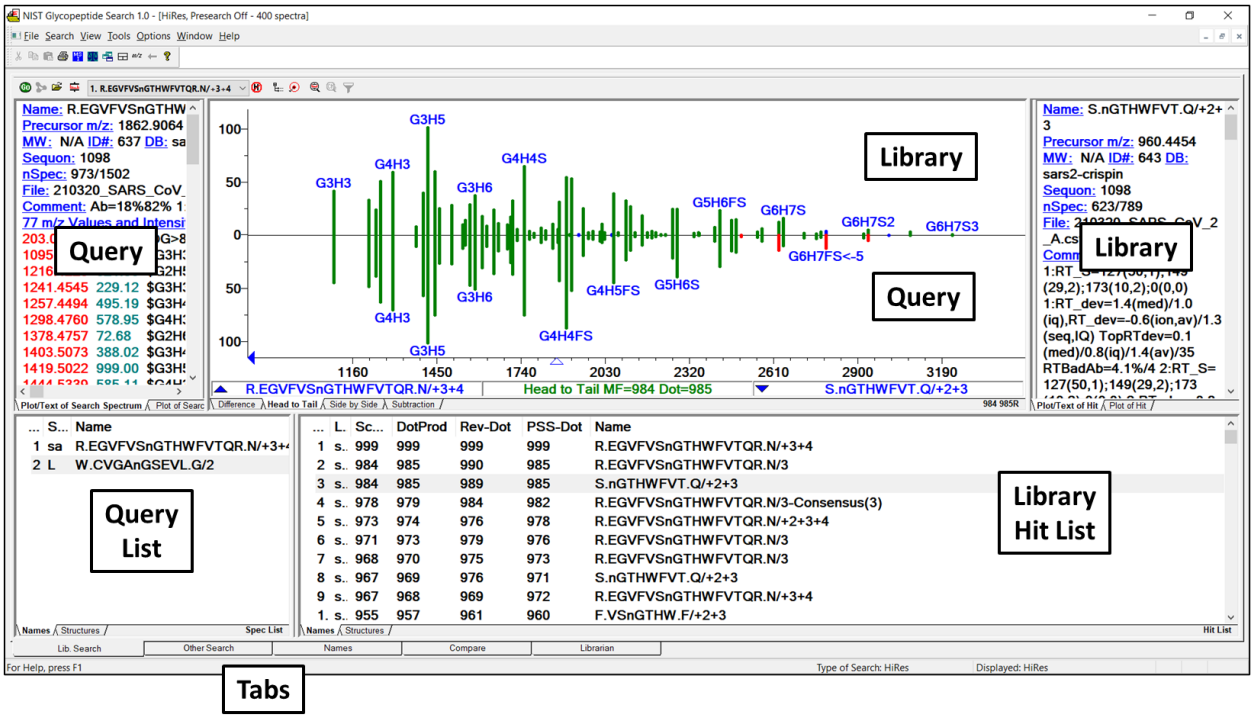
Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins

Concepcion A. Remoroza, Meghan C. Burke, Yi Liu, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Xiaoyu Yang, Stephen E. Stein
Mass Spectrometry Data Center, Biomolecular Measurement Division National Institute of Standards and Technology, 100 Bureau Drive Gaithersburg, MD 20899, US

Table of Contents

User Guide for NIST-MS Glycopeptide Abundance Distribution Spectra (GADS) libraries, software, and documentation.

User Guide for the NISTMS-GADS Program (Glycopeptide Search 1.0)



This document describes NISTMS-GADS.exe, an adapted version of NIST developed software that enables the processing of Glycopeptide Abundance Distribution Spectra (GADS). Only features required for GADS processing are described here. For an explanation of other features, many of which were designed for use with other varieties of mass spectral libraries, consult the online help system (F1) of documentation (includes pdf files in the folder containing this program: 'quick-start', 'tandem-library' and 'Ver24-man'). While this program contains controls for different types of spectra and libraries – as long as the settings

described here are used, none of those controls will affect GADS searches. Original GADS settings can be restored using the menu choice “File\Restore Settings” and then selecting “nistms-gads-default.ini”. This program is only intended for use with GADS libraries.

Technical details concerning the scoring and annotation discussed to in this document are given in the paper describing the method for extracting GADS from LC/MS mass spectral data of glycoprotein digests [submitted].

Five Views:

The NISTMS-GADS program provides 5 different views of the data, accessible through the 5 Tabs at the bottom of the screen shown in the screen display above.

Names Tab: Since this is the best way to start using this program, it is presented first. It enables the facile look up of individual GADS for exploring details of the data and its representation. Sorting is alphabetical and includes both the main ‘Name’ and any ‘synonyms’ for each GADS. Alternative names, ‘synonyms’, begin with ‘z’ to separate them from the main Name (sequence/charge), which is shown at the top of the text description of each GADS. GADS created from multiple charge states list these charges after the peptide sequence with charges separated by ‘+’. Consensus spectra for a single charge state are terminated with ‘Consensus(n)’, where n is the number of individual spectra combined to make the spectrum. A particularly useful ‘synonym’, begins with the letter ‘zs’, where all peptides that contain a specific ‘sequon’ are grouped together. A sequon is the asparagine amino acid onto which an N-glycan is attached and is represented as ‘n’ in the peptide sequence.

Lib. Search Tab: This set of five windows enables GADS library searching and comparison, showing the spectrum to be searched (the ‘query’ spectrum as the top plot) and the best library matches (the ‘hit list’, sorted by ‘scores, with the highlighted library spectrum shown as the bottom plot)’. The five windows are:

1. Spectrum list: This lists GADS that have been used for searching or sent there by selecting ‘Send to Spec. List’ from right mouse menu in any window. A double click performs a library search according to the settings in the library search menu described below. The most recently added spectrum is shown at the top of the list.
2. Hit list: Shows best matching library GADS sorted by score. It contains multiple columns with different score types, libraries, and peptides, each of which may be used for sorting by clicking on the column header.
3. Compare window: Shows a mirror (head-to-tail) plot of query GADS, shown as the upper plot with the text and plot to the left of the spectrum plot. It is also highlighted in the spectrum list. The library GADS is the bottom spectrum,, which is also shown to the right of the plot and highlighted in the hit list.
4. Query spectrum: Shows text/plot separated by a dividing bar of the query GADS used for searching and highlighted in the spec list.
5. Library spectrum: The ibrary spectrum shown at the bottom of the comparison window and highlighted in the hit list.

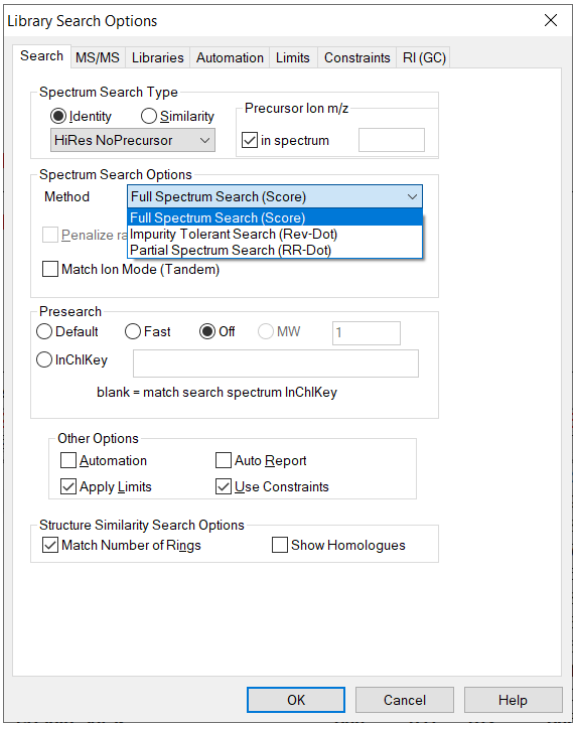
Spectrum Text Display (appears in same format in other Tabs along with plot):

On the Lib. Search Tab query and library spectra are displayed at the upper left and right, respectively, in both plot and text format, each separated by a vertical sliding bar. The text fields are described later in the Data Format section

Library Search Options Dialog (button on right below – near the top on *Lib. Search* tab):

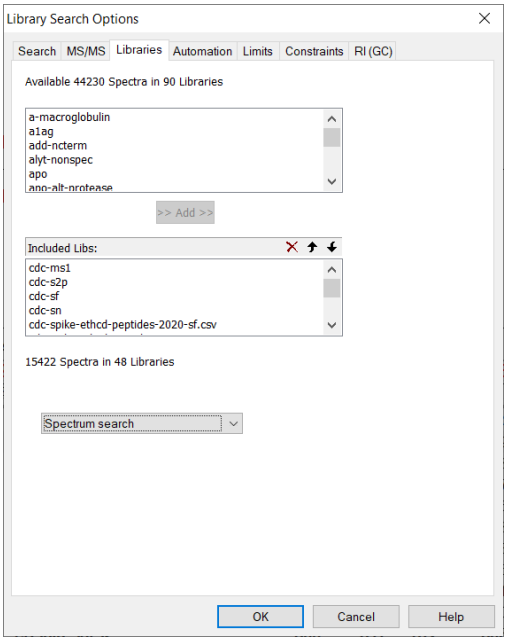
Search Tab: All search settings are available through the seven tabs in the complex dialog box below. In the first tab, three search types are possible, whose score serve to sort the hit list. The sequential search is recommended to ensure that all library GADS are compared, and the default is the 'high res no precursor' Identity search. Libraries for searching are selected in the Libraries tab.

By default, the 'Limits' button should be on, with minimum m/z set to 10 in that tab. Constraints can be useful for restricting searches (see later) but are not necessary. The other tabs and buttons are generally not relevant for GADS searches.

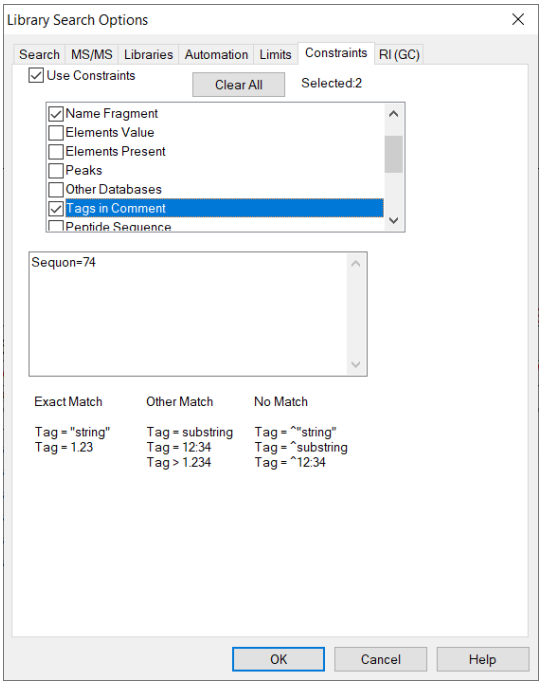


Two other Relevant tabs in the Library Search Options dialox box, *Libraries* and *Constraints*, are described below.

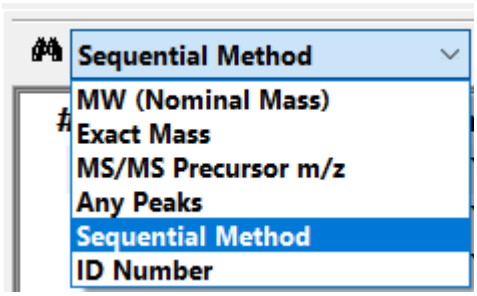
Libraries Tab: Lists all available libraries in the upper box. Double clicking on one transfers it to the lower box, which contains libraries used in the search. Double clicking on any of those in the lower box removes them from the search list. Only the 'Spectrum search' selection at the bottom is relevant for GADS searching.



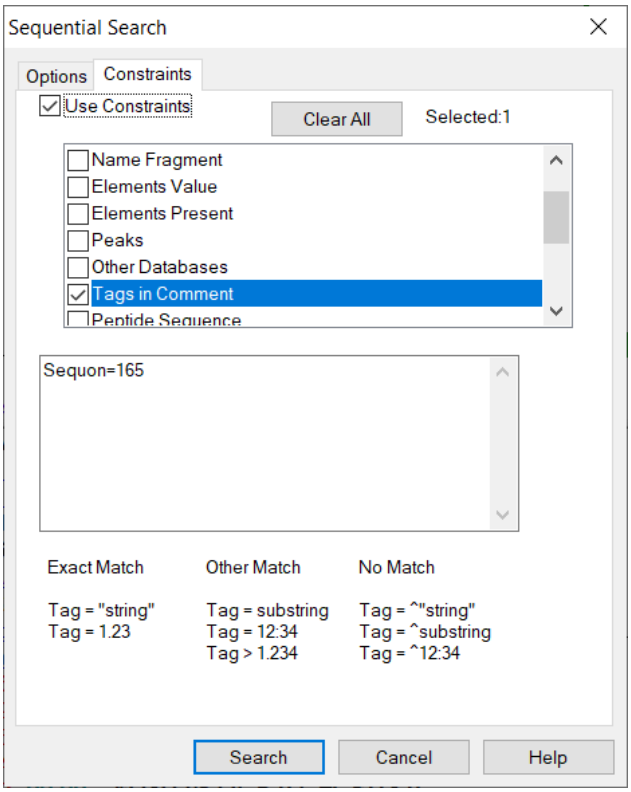
Constraints Tab (on the 'Library Search Options' Dialog): Two useful constraints for GADS are
1) 'Name Fragment', which specifies which characters must be in the full name of the glycopeptides (including the glycan). For example, entering /+ restricts GADS that combine multiple charge states or the letters 'NAT' would require a sequon containing this sequence in any ID.
2) 'Tags in Comment' can constrain searches using Tag=Variable values given in the comment field of the GADS (for example, Sequon=17, see later for pre-defined fields).
Additional information including how to perform more complex queries, can be accessed with the help button at the bottom.



Other Search For finding groups of GADS that meet certain criteria, such as Sequon, series of amino acid composition or tag=value information in the Comments field of the GADS. This is done using the 'Sequential Search' selection in the combo box at the top left (or the binocular button). None of the other selections are relevant for GADS searching.

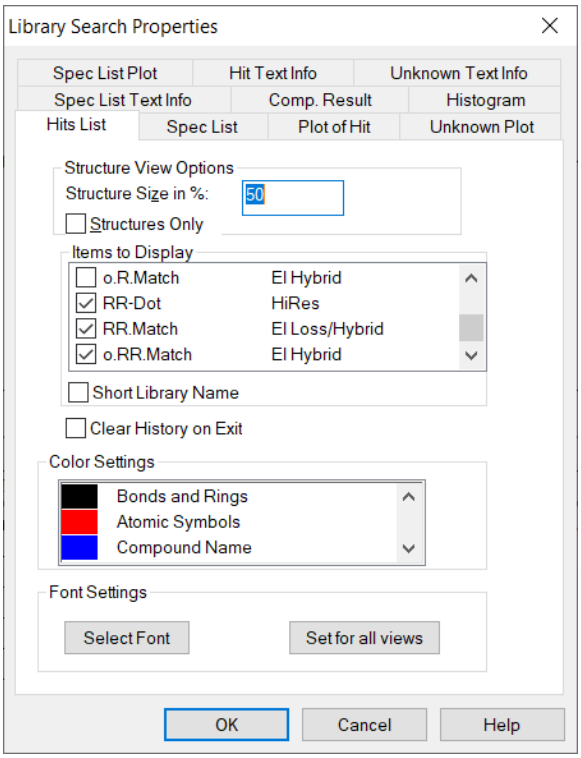


This invokes a dialog that allows the specification of libraries and search constraints (image below). Results of an example search for one library is shown on startup using the 'Sequon=' constraint. These constraints are also available in the Lib. Search dialog box (above)



Properties for Each Tab View

You can change many display features for all windows in each tab by selecting ‘Properties’ from the right mouse button list. Below are the Properties that appear for the ‘Lib. Search’ Tab. Each region of the screen is associated with a different tab. The image below shows the controls for the ‘Hits List’ section, where you can select various columns and colors, as well as the font for just the current window. These selections depend on the particular Tab view that was active when you selected Properties. Note that the Help button provides more information.



DATA FORMAT

GADS Text Format: Text underlying displayed GADS are shown and are in an expanded '.msp' format used for many years by NIST Mass Spectral data programs. They are serve for input and output in as illustrated below (highlighted fields are required for input GADS).

Name: W.CVGAnGSEVL.G/2
PrecursorMZ: 1004.4597
MW: Not used, *DB*: Name of library
Synon: zs 0076 W.CVGAnGSEVL.G/2
Comments: Sequon=76 File=Tg_02_RG_NL_2021-03-30_Tg-C1A2-0_380-2000-120_HCD40_350min_2ug_Chymo-AspN_I_Pos.csv
nSpec=123/200 Protease=T log(MaxAb)=9.1 Totab/Maxab=8.45 nGoodPks=41/41 Max2Med=11.09 GoodAb=1.00 nGoodPks=5/5 Qual=4.2 GoodAb=0.96
Num Peaks: 6
1216.4229 949.35 "$G2H5/s442,#4,r51.5,p2.1"
1581.5551 331.37 "$G3H6/s394,#3,r51.1,p-0.2"
1622.5816 199.80 "$G4H5/s398,#2,r51.3,p1.1"
1751.6242 246.95 "@G4H4S>19/s58,#2,r78.9,p0.5"
1768.6395 999.00 "$G4H5F/s566,#5,r51.6,p2.2"
2059.7349 500.10 "$G4H5FS/s461,#3,r59.7,p-1.2"

Descriptions of this data follows:

Name: Peptide sequence for the GADS, followed by /+charge(s) (or Consensus(n))
PrecursorMZ: This is the peptide mass – MZ is used for compatibility with mass spectral searching. Not used in searching.
MW: Not used, *DB*: Name of library
Synon: This is listed along with 'Name' in the Name tab, sorted alphabetically
Selected Values: Taken from comment field below with <tag>=<value> as specified in *Options/Comment Field Display* dialog box.
Comments: Various information for the GADS, given in <Tag>=<Value> format.Fields used for NIST GADS libraries are presented below.
Sequon= sequence number of asparagine attached to glycan
File= name of original data file
Class= shows fractional abundance of 5 categories of glycans in the GADS, including HiMannose (HiMan), Hybrid, Complex, Fucose, and Sialyl (HiMan=G2H5-9, Hybrid=G3H5-8, Complex=G4-6H5-7, Hybrid and Complex may also contain Fucose and Sialyl).
nSpec= number of identified MS2 spectra, good score/all scores
Protease=T (trypsin), C(chymotrypsin),G(gluc),A(alphalytic)
log(MaxAb)= log10 of XIC (MS1) abundance
TotAb/MaxAb= sum of abundances/maximum (base peak) abundance
nRT:OK/notOK(>4)=number of peaks withinin retention time tolerace/outside of tolerance
nGoodPeaks= number of good scoring/all glycan peaks
Max2Med= maximum abundance/median abundance
Qual= GADS quality = nGoodGlycans*(nGoodGlycans/nAllGlycans)

Unoccupied= if found, percent of the non-glycosylated abundance of this sequence
Note that any of these tags=value representations may be shown as separate lines in the text display and in the figure title in the order they are entered in to Options/Comment Field Display box in the main menu.

Good implies Byonic score > 30 and retention within tolerance of 4 minutes. May change in future versions.

Num Peaks: Required – number must match the number of mass/abundance lines that follow *Mass Abundance Pairs"</ Peak Annotation>"* . Each glycan peak is represented as a pair of mass/abundance pairs optionally followed by a string of characters in quotes. This string has two components separated by a '/'. The first component is displayed above the peak, and if it begins with a special symbol it is assigned a selected color (specified in the Properties dialog for that plot). These special symbols, their default colors and their mapping to colors selectable in the Properties dialog are:

Symbol	*Default*	*Properties Setting*
$	Green	Peptide Peaks
&	Blue	Y Peaks
@	Red	Special Peaks
|	Black	none
other	Red	Peaks

For each GADS, thes space-separated text information can be viewed in text spectrum display window as described earlier in "Comments Field".

Following each mass/abundance value and glycan display text, a number of quantities are presented for the identified glycopeptide. These comma-separated values follow the forward slash ('/'):

s<value>	value=Byonic score
p<value>	value=ppm deviation for highest scoring glycopeptide
#<integer>	number of MS2 identification above score of 30
r<value>	retention time in minutes for XIC maximum
nr<integer>	number of replicate spectra with this peak (consensus spectrum only)
+n+m..%..%	percent abundance from charge +n and +m (combined charge spectrum only)

GADS Libraries:

Most GADS libraries distributed with this software contains data for a single protein from a single lab (several libraries contain 2 closely related proteins). Each library appears as a subfolder in the folder that contains the NISTMS-GADS.exe program. These libraries is described in Table I of the paper that describes the GADS concept.

The program LIB2NIST is also provided with this software. It will convert a file in .msp format described above into a searchable library file. Examples of the .msp format for any GADS may be generated by

selecting ‘export’ from the right mouse menu. It can also convert a GADS lirary to a single .msp file. Also, libraries may be added to or deleted using the **Librarian Tab**. This tab view also allows editing of GADS and saving in a library or in the temporary “Spec.List” shown on the **Lib. Search Tab**. The file/open menu choice will read GADS files in .msp (or .mspec, which is the same format) that have been exported previously or prepared by the user.