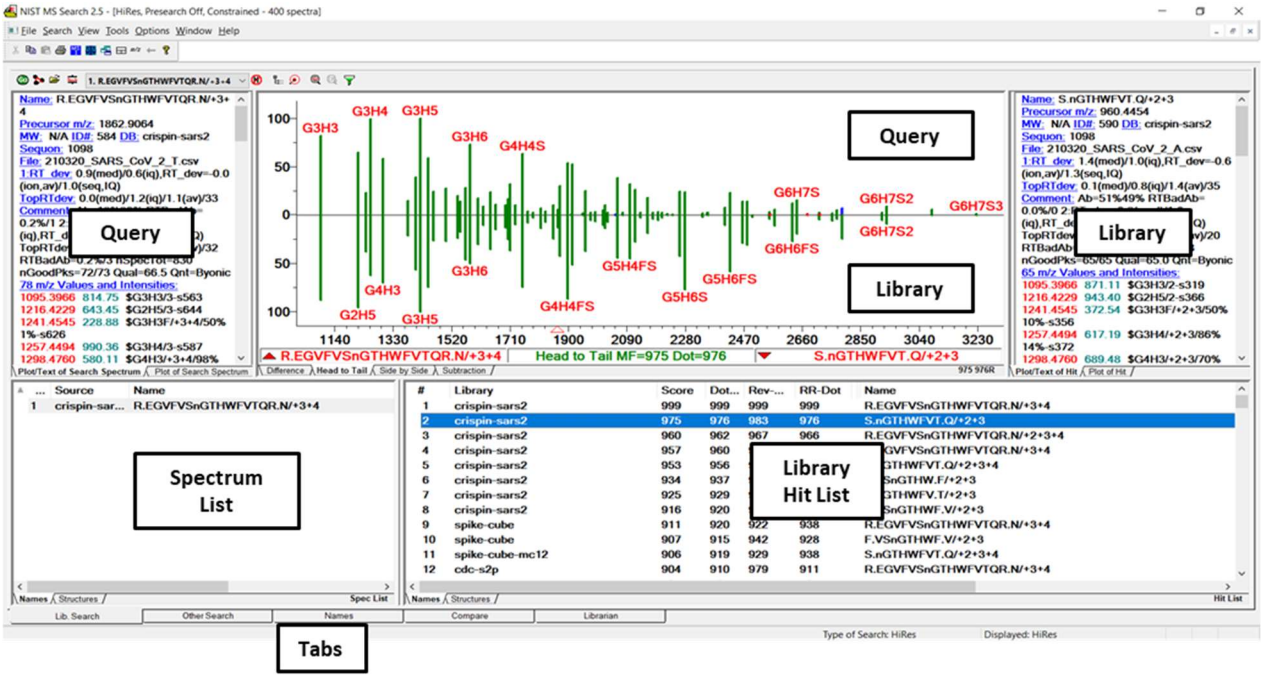
User Guide for the NISTMS-GADS-Glycopeptide Program
(Glycopeptide Search 1.0)

Contents

Guide to Program Operation



This document describes an adapted version of NIST developed software for mass spectral library analysis that enables the processing of both Glycopeptide Abundance Distribution Spectra (GADS) and its underlying individual glycopeptide ion mass spectra. These are implemented in the program NISTMS-GADS.exe, which is titled Glycopeptide Search 1.0. This program has five distinct views, one of which, the Library Search view, is shown above with individual regions labeled.

Only features required for GADS and glycopeptide processing are described in this document. For an explanation of other features, many of which were designed for use with other types of mass spectral libraries, consult the online help system (F1) or accompanying documentation (includes pdf files in the folder containing this program: ‘quick-start’, ‘tandem-library’ and ‘Ver24-man’). While this program contains controls for a variety of spectra and libraries – as long as the settings described here are used, none of those controls will affect GADS or glycopeptide spectra display or searching. Original settings can be restored using the menu choice “File\Restore Settings” and then selecting “nistms-gads-default.ini”.

Technical details concerning the scoring and annotation of GADS discussed to in this reference: Remoroza, Concepcion A., Meghan C. Burke, Yi Liu, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Xiaoyu Yang, and Stephen E. Stein. "*Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins*." Journal of Proteome Research 20, no. 9 (2021): 4475-4486.

Glycan and glycopeptide annotation is described in: Yang, Xiaoyu, Pedatsur Neta, Yuri A. Mirokhin, Dmitrii V. Tchekhovskoi, Concepcion A. Remoroza, Meghan C. Burke, Yuxue Liang, Sanford P. Markey, and Stephen E. Stein. "*MS_Piano: A Software Tool for Annotating Peaks in CID Tandem Mass Spectra of Peptides and N-Glycopeptides*." Journal of Proteome Research 20, no. 9 (2021): 4603-4609.

Five Views (main tabs at bottom):

This program provides 5 different views of the data, accessible through the 5 Tabs at the bottom of the screen as shown in the screen display above.

Names Tab (third main tab): Since this is the best way to examine spectra, it is presented first. It enables the facile look up and examination of individual GADS or glycopeptide spectra. Basic sorting is alphabetical by peptide 'Name', including charge(s) and for glycopeptide spectra also glycan modification. Alternative sorting employs prefixes followed by the 'Name'. GADS can also be sorted by sequon – these begin with prefix 'zs<sequon>' where sequon is the numerical position of the glycan attachment to the protein. This groups together GADS for the same sequon. Glycopeptide spectra are sorted in three ways, (1) by glycan ('xg_' prefix); (2) 'Scan_<file number>_<Scan number>' prefix; (3) alternative names ("Alt_id"_ prefix). Some GADS spectra are the sum of GADS of multiple charge states which have with charges separated by '+'. Consensus GADS for a single charge state are terminated with 'Consensus(n)', where n is the number of individual spectra combined to make the spectrum. Only in the last case are GADS derived from more than a single LC/MS run. Each of the glycopeptides represented by peaks in GADS have a corresponding entry in the glycopeptide library and show the name of the underlying data file and scan number. GADS libraries have _GADS affixed to their name while glycopeptide libraries do not.

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

1. Query list: This lists GADS or glycopeptide spectra that have been used for searching, imported, or sent there by selecting 'Send to Spec. List' from right mouse menu in any window. A double click in the list performs a library search according to the settings in the library search menu described later. The most recently added new spectrum is shown at the top of the list. The spectrum highlighted in this list is shown as the upper spectrum in the plot (with text to the left in the default view)
2. Hit list: Shows best matching library spectra sorted initially 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. Information shown differs between GADS and glycopeptide ion searches (see below), which is selectable from Properties Dialog or hiding columns using the list divider bar at the top of the list.
3. Spectrum comparison plot: Shows a mirror (head-to-tail) plot of the query spectrum (GADS or glycopeptide), and selected library spectrum. The query spectrum is upper spectrum with the text and plot to the left of the spectrum plot. The library spectrum is at the bottom with individual spectrum information at the right. The query spectrum is highlighted in the Query list and the library spectrum highlighted in the hit list.
4. Query spectrum: Shows text/plot separated by a dividing bar of the query spectra used for searching and highlighted in the spec list.
5. Library spectrum: The library spectrum shown at the bottom of the comparison window and highlighted in the hit list.

Note that the above description applies to the default layout and may change if other layouts are selected using rectangle on the button bar at the top of the screen. This as well as resizing windows allows the users to customize the screen which can be saved in File/Save Configuration menu choice.

Hit List (item 2 above) Examples

Examples of Glycopeptides and GADS Hit Lists
Can select fields for display using Properties Dialog Box or can hiding columns by moving column divider at top, to the left.

Glycopeptide Identifications – shows all tandem spectrum information

#	Library	Score	DotProd	Rev-Dot	NumMP	PSS-...	Prec. Type	Instr...	E...	Mods	Name
1	hlf_nist	999	999	999	193	999	[M+3H]3+	HCD	15	G:G3H4S CAM	TAGWNIPMGLLFNQTGSCKFDE/3_2(12,N,G:G
2	hlf_nist	931	949	956	183	979	[M+3H]3+	HCD	15	G:G3H4S CAM	TAGWNIPMGLLFNQTGSCKFDE/3_2(12,N,G:G
3	hlf_nist	928	944	957	171	970	[M+3H]3+	HCD	15	G:G3H4S CAM	TAGWNIPMGLLFNQTGSCKFDE/3_2(12,N,G:G
4	hlf_nist	911	938	947	179	977	[M+3H]3+	HCD	15	G:G3H4S CAM	TAGWNIPMGLLFNQTGSCKFDE/3_2(12,N,G:G
5	hlf_nist	851	885	897	172	922	[M+3H]3+	HCD	15	G:G4H5SNa CAM	TAGWNIPMGLLFNQTGSCK/3_2(12,N,G:G4H5
6	hlf_nist	851	877	911	137	907	[M+3H]3+	HCD	15	G:G4H5SNa CAM	TAGWNIPMGLLFNQTGSCK/3_2(12,N,G:G4H5
7	hlf_nist	850	875	923	126	899	[M+3H]3+	HCD	15	G:G4H5SNa CAM	TAGWNIPMGLLFNQTGSCK/3_2(12,N,G:G4H5

GADS Identifications - fewer columns needed

#	Library	Score	DotProd	Re...	Num...	PS...	Name
1	hlf-nist-...	999	999	999	14	999	TAGWNIPMGLLFnQTGSCK/+2+3
2	hlf-nist-...	998	998	998	13	998	TAGWNIPMGLLFnQTGSCK/3
3	hlf-nist-...	945	964	964	14	986	TAGWNIPMGLLFnQTGSCK/3
4	hlf-nist-...	941	963	963	14	984	TAGWNIPMGLLFnQTGSCK/+3+4
5	hlf-nist-...	937	945	965	6	945	TAGWNIPMGLLFnQTGSCK/4

Spectrum Text Display (items 4 and 5 above, spectrum hidden) appears in same format in various Tab Views along next to the spectrum plot). Glycopeptide example at left, GADS at right below

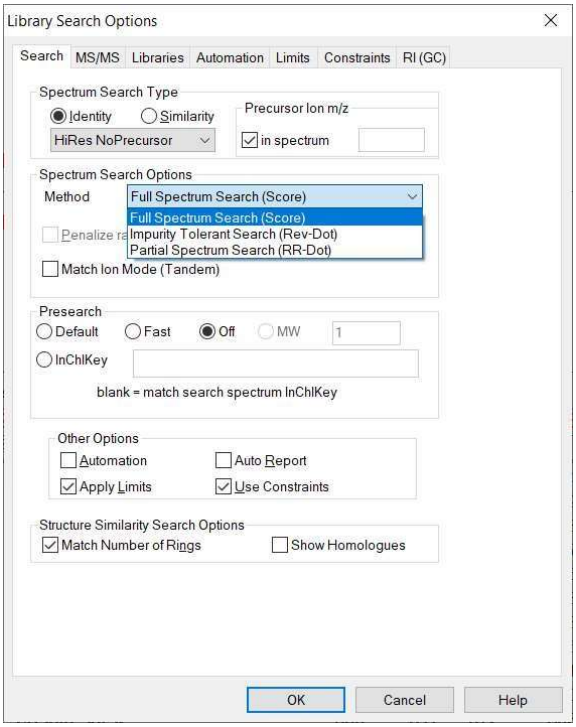
Name: DLCGCYSVSSVLPGCAEPWNHGK/4_4(2,C,CAM)(4,C,CAM)(9,S,G:G3H3S2)(14,C,CAM)
Precursor type: [M+4H]4+
Instrument type: IT-FT/ion trap with FTMS
Collision energy: 30%
Precursor m/z: 1068.4353
MW: N/A ID#: 279 DB: 2021-09-09_igha_o-glycans_byonic_hcd_cid_good
Sequon: 187
Fm_pm: 0.0038
Purity: 55,91
Score: 139.14
Theo_mz_diff: 1.2ppm
Scan: 01_28548
Seq_annotated: 0000000000 1100000000 010
RT: 79.1961
MGF: IgA_01_urea_L1_2021-09-09_380-2000_120_HS152535CIT_150min_300ng_TrypLysC_I_Pos.raw.mgf

Name: EGVFVSnGTHWFVTQRNFY/3
Precursor m/z: 2287.081
MW: N/A ID#: 2 DB: History
Sequon: 1137
nSpec: 76/81
File: SP-IBBR-UMD_04_RG-in-solution_NL_2022-02-22_380-2000_120_HS152535CIT_230min_1ug_AspN_S3_I_Pos.csv
RT_S: 115(14,18);122(10,18);128(1,0);135(0,0);142(0,0);
Protease: N
TopRTdev: 0.0(med)/0.1(iq)/0.2(av)/8/
nScore:Hi/Lo: 23/0
nRT:OK/notOK(>4): 23/0
Comment: FullName=R.EGVFVSnGTHWFVTQRNFY.E Protein=sp|P0DTC2|IBBR-MD_Spike RTdev_S=0.2(med)/0.1(iq),RT_dev=-0.0(ion,av)/0.3(seq,IQ) RTBadAb=0.0%/0 SDelay=6.79 RTDevMed=0.74 AllowDev=4.47

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 sections

Library Search Options Dialog (button on right below – buttons at top of Lib. Search tab):

This brings up a complex dialog box with 7 tabs across the top.

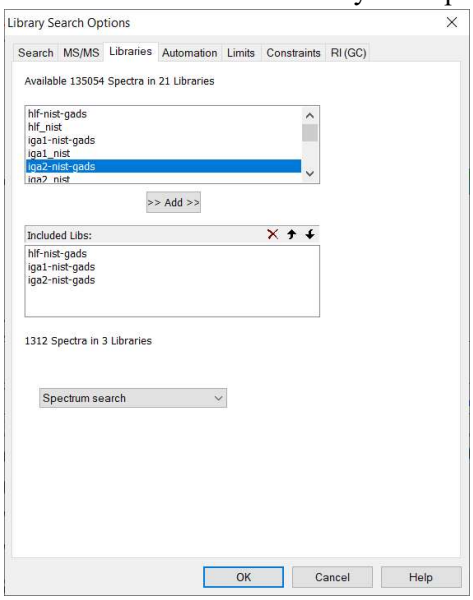


Relevant tabs for GADS and glycopeptide spectra a described below.

Search Tab: Use the above selections for typical GADS or glycopeptide search. The default scoring type is the 'high res no precursor' Identity search. Scores for all three score types are displayed by default in the hit list, which may be sorted by those alternate scores later. Libraries for searching are selected in the Libraries tab. Presearch may be set to 'Default' for faster search, but may miss spectra that do not match the query well.

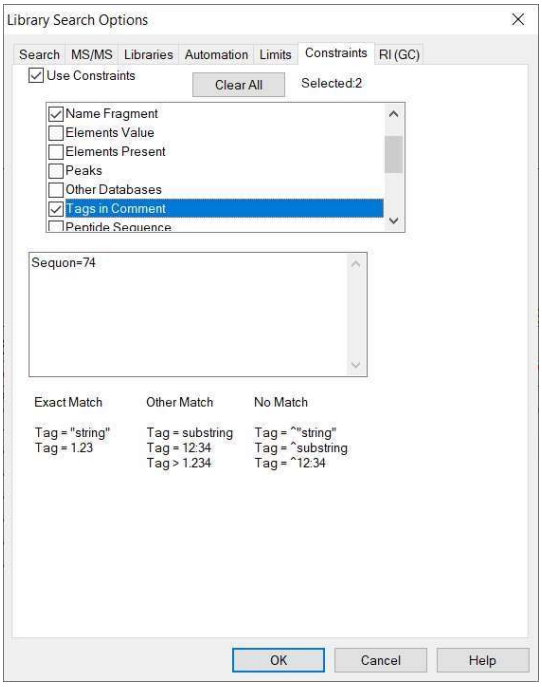
The 'Apply Limits' button should be on, with the 'Limits' tab minimum m/z set to a value less than your spectra, for example 10. 'Use Constraints' can be useful for restricting searches (see later) according to annotation in each spectrum but are not necessary. The other tabs and buttons are generally not relevant for GADS or glycopeptide searches.

Libraries Tab: A list of all available libraries is shown in the upper box. Double clicking on one transfers it to the lower box, which contain libraries used in the search. Double clicking on any of those in the lower box removes them. Only the ‘Spectrum search’ selection at the bottom is relevant.

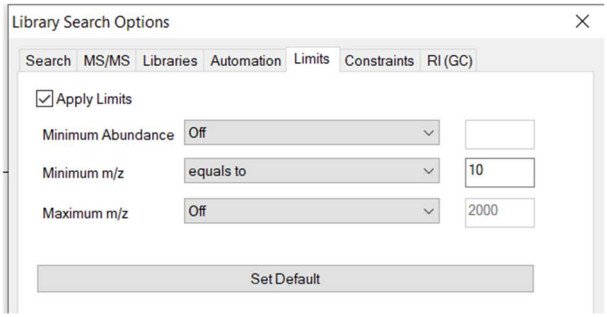


Note that GADS libraries end in ‘-gass’ – but this is not necessary and mixing library type will do no harm but only search relevant libraries.

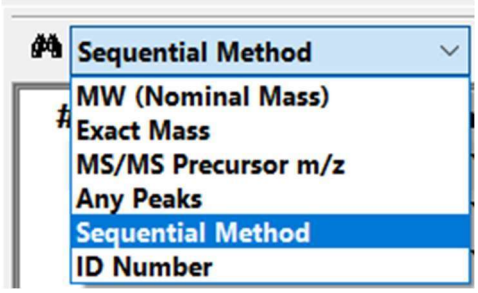
Constraints Tab (on the ‘Library Search Options’ Dialog): Two useful constraints for GADS are
1) ‘Name Fragment’, specifies which text must be in the full name of the glycopeptides (including the glycan). For example, entering /+ restricts the hit list to GADS that combine multiple charge states, or, for example the letters ‘NAT’ would require a sequon containing this amino acid sequence in each ID.
2) ‘Tags in Comment’ can constrain searches using Tag=Variable values given in the comment field (for example, Sequon=17, is applicable to GADS, see later for pre-defined fields). More complex cases can be handled. For guidance press the help button at the bottom.



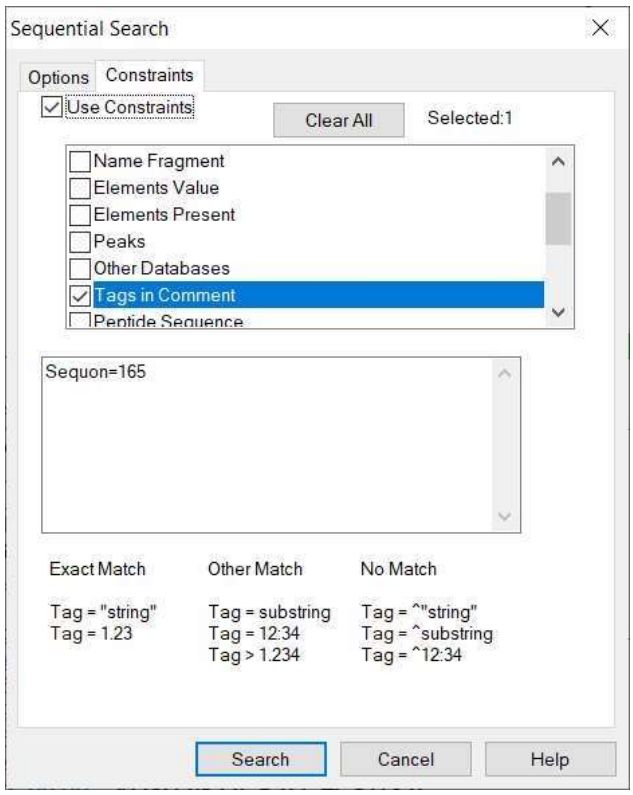
The Limits tab should have a low value set for minimum m/z – other values are optional



Other Search Tab (second main tab): For finding groups of spectra that meet certain criteria, such as Sequon, series of amino acid composition or tag=value information in the Comments field of the spectra. 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 or glycopeptide searching.

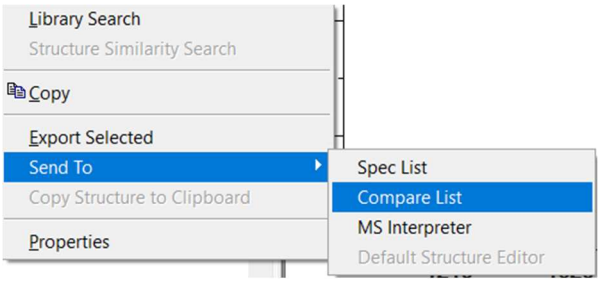


This invokes a dialog that allow the specification of libraries and search constraints (image below). Results of an example search for one library is shown on startup using the 'Sequon=' constraint. These constraints are also available in the Lib. Search dialog box (above)



Compare Tab (fourth main tab)

This enables comparison of any spectra sent to the compare window by selecting 'Compare List' from the menu generated by right clicking on a spectrum.



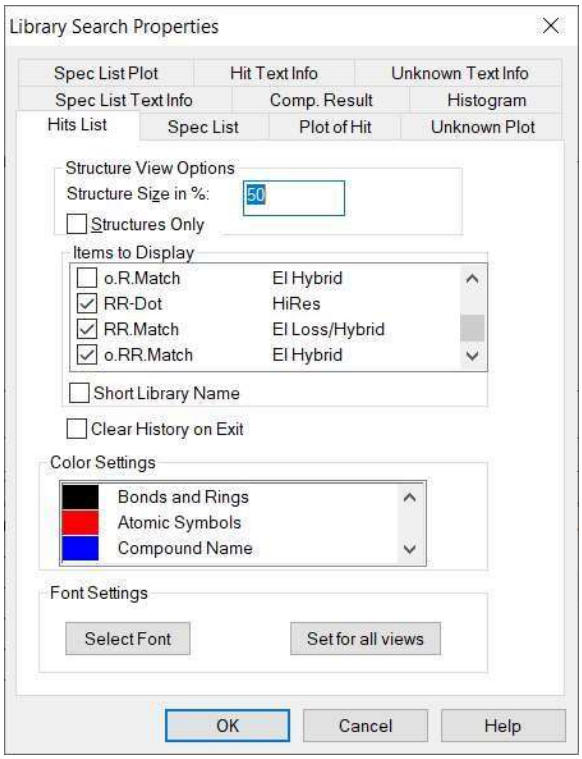
The behavior is described in main program documentation, which can be modified by selecting properties (see below). The maximum number of spectra displayed is specified in the Properties Dialog (select from right mouse menu)

Libraries Tab (fifth main tab)

This displays the same spectra that are in the 'spec list' (above) enabling the editing of spectra in a library and creating new libraries. See main documentation for details.

Properties Dialog for Each Tab View

You can change many display features in the active window windows by selecting 'Properties' from the right mouse button list. Choice depend on the specific window selected. Below are the Properties that appear for the 'Lib. Search' Tab, which is the most complex. Each window in the View is represented as a tab in this Dialog Box. Shown below are Properties for the 'Hits List' section, where you can select various columns and colors, as well as the font for just the current window or all windows on the screen. These selections depend on the particular Tab view that was active when you selected Properties. The Help button provides more details.



Glycan Representation

Glycan compositions are represented using the following symbols for individual components:

G = GlcNac (or GalNac)
H = Hexose (Mannose or Galactose)
F = Fucose
S = Sialyl (NeuAc)
Sg = NeuGc
So = Sulfation
Po = Phosphorylation

Composition of a given glycan is represented in a manner similar to that of a chemical formula, with components in the order of the above list. For example,

G4H5FS corresponds to 4 GlcNac, 5 Hexose, 1 Fucose, and 1 Sialyl group

GADS 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. An example is shown below followed by a description of each data type:

Name: W.CVGAnGSEVL.G/2
PrecursorMZ: 1004.4597
MW: Not used, DB: Name of library
Synon: zs 0076 W.CVGAnGSEVL.G/2
Class: HiMan=0.005,Hybrid=0.000,Complex=0.977,Fucose=0.819,Sialyl=0.464
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 each data type follows:

Name: Peptide sequence for the GADS, followed by /+charge(s) (or Consensus(n))
PrecursorMZ: This is the mass of the peptide only – The MZ field is for compatibility with mass spectral searching. It is not used for GADS.
MW: Not used, DB: Name of library
Class: Relative abundances of different classes of glycans, described below
Synon: This is listed along with 'Name' in the Name tab, sorted alphabetically
Selected Values: Extracted from comment field below with <tag>=<value> as specified in Options/Comment Field Display dialog box. Users may select their order and identity as they wish.
Comments: Various information for the GADS, given in <Tag>=<Value> format. Specific fields used for NIST GADS libraries are presented below. Some are use primarily for NIST quality control.
Sequon= sequence number of asparagine attached to glycan
File= name of original data file – may have date and other data
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 within retention time tolerance/outside of limit
nGoodPeaks= number of good scoring/all glycan peaks
Max2Med= maximum abundance/median abundance
Qual= GADS quality = nGoodGlycans*(nGoodGlycans/nAllGlycans)
TotalUnoccupied= if found, percent of the abundance of this sequence without attached glycans

Good implies Byonic score > 30 and retention within tolerance (typically 4 minutes).

Num Peaks:<integer> This is the number of peaks - Required – must match the number of m/z : abundance lines that follow. Each peak has the format:

m/z Abundance "</Peak Annotation>"
Each of these glycan peaks is represented as pair of mz/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	Meaning	Properties Setting
$	Green	Good Peaks	Peptide Peaks
&	Blue	Uncertain	Y Peaks
@	Red	Bad	Special Peaks
|	Black	No Label	none
?, other	Red	Unknown	Peaks

For each GADS, a variety of space-separated textual information can be viewed in the "Comments" as described earlier.

Following each mass/abundance pair and glycan display information, a number of quantities are presented for each 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)

Glycan Class: Fractional abundance of each of five classes of glycans:
HiMan : high mannose, G2Hn, n from 5 to 9
Hybrid: G3Hn, n from 5 to 8, with any number other types, F, S, ..
Complex:, G>3Hn, n from 3 to 7, with any number other types, F, S, ..
Fucosyl: glycans containing one or more fucose
Sialyl: glycans containing one or more sialyl groups

GADS Libraries

Most GADS libraries distributed with this software contains data for a single protein from a single source (some contain multiple proteins). Each library appears as a subfolder in the folder containing the NISTMS-GADS.exe program. These libraries described the paper associated with this program.

The program LIB2NIST is also provided with this software. It will convert a file in .msp format described above into a searchable library file or convert a library file to .msp format. Examples of the .msp format for any GADS may be generated by selecting 'export' from the right mouse menu. 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 treated as the same format) that have been exported previously or prepared by the user.

Glycopeptide Data Format

Glycopeptide Text Format: Text underlying displayed glycopeptides are similar in format to those in the NIST Peptide and Tandem libraries. They are displayed, imported, and exported in an expanded '.msp' format used for many years by NIST Mass Spectral data programs. The Comments field contains a number of glycopeptide specific quantities in <tag>=<value> format. The principal values are shown below and the full set, which includes information used by NIST for quality control, is given in the file "glycopep-annotation.txt". Each identified peak also has annotation unique to glycopeptide product ions using the format described above in the Glycan Representation section. An example the text follows:

Name: TAGWNIPMGLLFNQTGSCK/3_2(12,N,G:G4H5S)(17,C,CAM)
PrecursorMZ: 1336.9041
Precursor_type: [M+3H]3+
Ion_mode: P
Collision_energy: 15,25,35%
Instrument_type: HCD
Ionization: ESI
Spectrum_type: MS2
Synon: Scan_04_44524_TAGWNIPMGLLFNQTGSCK/3_2(12,N,G:G4H5S)(17,C,CAM)
Synon: xG_G4H5S_44524_04_TAGWNIPMGLLFNQTGSCK/3
Synon: z02030A_TAGWNIPMGLLFNQTGSCK/32(12,N,G:G4H5S)(17,C,CAM)
DB#: 731
Comment:
<a number of <tag>=<value> pairs are contained in this text field. Tags of most interest to users are described below using typical <values>, followed by a brief explanation after a '//' delimiter – for a full description which includes information used quality control, see file glycopep_annotation.txt *in the same folder as this file,*
Parent=1336.9041 // precursor m/z
Mods=2(12,N,G:G4H5S)(17,C,CAM) // modifications, in format n(m,AA, type of modification) where n=number of modifications, m=position of modification in peptide sequence, counting from 0 at N-terminus, AA=amino acid point of attachment, type of modification, if glycan : G:glycan, CAM is carbamidomethyl.

NCE=15,25,35 // energy setting – stepped HCD in this case
MGF=hLf_13_urea_L1_2021-07-02_Lf-TL1-0_380-2000_120_HS152535CIT_150min_2ug_TrypLysC_I_Pos.raw.mgf // input data = contains protease, date of measurement and other information

Scan=04_44524 // 4th input file, scan 44524 – 04 denotes file using an arbitrary integer
Score=443.09 // Byonic score
RT=99.3949 // retention time for underlying extracted ion chromatogram
Full_name=R.TAGWNIPMGLLFNQTGSCK.F // sequence with flanking amino acids
Protein=">sp|P02788|TRFL_HUMAN Lactotransferrin OS=Homo sapiens" // protein name from fasta file
Sequon=497 // glycan position in protein sequence
Num Peaks: 203 // number of peaks – this must be followed 203 peaks, each in a separate line

Part of spectral output : m/z abundance annotation absolute format = relative to precursor (if appropriate) followed by ppm error – note that '^' precedes charge state
…
1661.4301 11.46 "?"
1676.7372 317.83 "Y0+{G3H4}^2=p-{GHS}^2/2.3ppm"
1677.2405 500.83 "Y0+{G3H4}^2+i=p-{GHS}^2+i/3.4ppm"
1677.7411 340.03 "Y0+{G3H4}^2+2i=p-{GHS}^2+2i/3.1ppm"
1678.2399 33.33 "Y0+{G3H4}^2+3i=p-{GHS}^2+3i/1.7ppm"
1733.7911 13.77 "?"
1741.2428 19.11 "Y0+{G3H3S}^2=p-{GH2}^2/-6.8ppm"
1741.7731 26.68 "Y0+{G3H3S}^2+i=p-{GH2}^2+i/9.8ppm"
1742.2621 21.48 "Y0+{G3H3S}^2+2i=p-{GH2}^2+2i/2.8ppm"
1778.2759 33.98 "Y0+{G4H4}^2=p-{HS}^2/1.6ppm"
1778.7760 27.39 "Y0+{G4H4}^2+i=p-{HS}^2+i/0.8ppm"
1822.2892 42.15 "Y0+{G3H4S}^2=p-{GH}^2/4.5ppm"
1822.7874 130.48 "Y0+{G3H4S}^2+i=p-{GH}^2+i/2.7ppm"
1823.2917 52.43 "Y0+{G3H4S}^2+2i=p-{GH}^2+2i/4.4ppm"
1858.8422 17.02 "?"
..

Descriptions of this data follows:

Name: Peptide sequence, followed by charge and modification (glycan)
PrecursorMZ: This is the glycopeptide mass (precursor ion)
MW: Not used, DB: Name of library
Synon: Alternate names for this precursor peptide as found 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: Metadata relevant to the spectrum given in <Tag>=<Value> format. Fields used for NIST glycopeptide spectra are described above and in the accompanying file: "glycopep_annotation.txt".