Characterization of Carcino Embryonic Antigen (CEA) Microheterogeneity: A Case Study Considering N-linked Glycosylation Forms

Ravi Amunugama, Michael Ford, Richard Jones and David Allen

MS Bioworks, LLC. 3950 Varsity Drive Ann Arbor, MI 48108

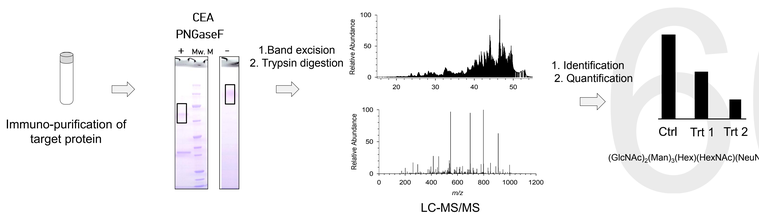
1. Abstract and introduction

Colorectal cancer (CRC) is the fourth leading cause of cancer related deaths in the United States[1]. A critical component of CRC care is post-surgical monitoring for cancer recurrence. CEA is a tumor marker for the clinical management of CRC which has the specific utility of monitoring post-operative disease recurrence. After surgery to remove cancerous tissue, the level of CEA in blood can be periodically monitored using an immunoassay. If the levels begin to rise above 6.0 ng/mL there is a high correlation with recurrence of the cancer[2].

CEA, like many tumor markers is a glycoprotein and there is a significant body of work showing protein glycosylation is greatly affected by diseases such as cancer. We are investigating glycosylation forms of CEA as sensitive and specific biomarkers of CRC. This presentation outlines our methods and includes qualitative and quantitative CEA glycosylation form data.

1. *Cancer Trends Progress Report – 2009/2010 Update*, National Cancer Institute, NIH, DHHS, Bethesda, MD, April 2010, http://progressreport.cancer.gov.
2. M.G. Fakih and A. Padmanabhan "CEA Monitoring in Colorectal Cancer" Oncology 20, 2006, 1- 15

2. Technology Overview



- 1µg of purified protein (with and without treating with PNGaseF) was separated on a 10% Bis-Tris NuPage gel (Invitrogen) in the MES buffer system.
- The segments corresponding to the modified and unmodified proteins were excised and processed by in-gel digestion using a robot (ProGest, DigiLab) and sequencing grade trypsin (Promega).
- Each digest is processed by LC-MS/MS using a reverse phase gradient.

3. Glycosylated Sites Mapping

IPI00880101: Gene-Symbol = CEACAM5 Protein Amino Acids Sequence (IPI Human Database)

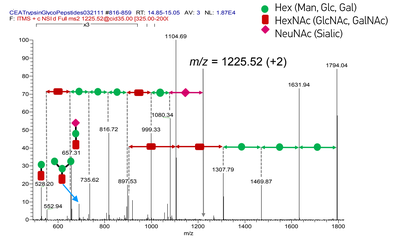
```
  1 XSAPPHRWCI PWQRLLLTAS LLTFWNPPTT AKLTIESTPF NVAEGKEVLL
 51 LVHNLPQHLF GYSWYKGERV DGNRQIIGYV IGTQQATPGP AYSGREIIYP
101 NASLLIQNII QNDTGFYTLH VIKSDLVNEE ATGQFRVYPE LPKPSISSNN
151 SKPVEDKDAV AFTCEPETQD ATYLWWVNNQ SLPVSPRLQL SNGNRTLTLF
201 NVTRNDTASY KCETQNPVSA RRSDSVILNV LYGPDAPTIS PLNTSYRSGE
251 NLNLSCHAAS NPPAQYSWFV NGTFQQSTQE LFIPNITVNN SGSYTCQAHN
301 SDTGLNRTTV TTITVYEPPK PFITSNNSNP VEDEDAVALT CEPEIQNTTY
351 LWWVNNQSLP VSPRLQLSND NRTLTLLSVT RNDVGPYECG IQNKLSVDHS
401 DPVILNVLYG PDDPTISPSY TYYRPGVNLS LSCHAASNPP AQYSWLIDGN
451 IQQHTQELFI SNITEKNSGL YTCQANNSAS GHSRTTVKTI TVSAELPKPS
501 ISSNNSKPVE DKDAVAFTCE PEAQNTTYLW WVNGQSLPVS PRLQLSNGNR
551 TLTLFNVTRN DARAYVCGIQ NSVSANRSDP VTLDVLYGPD TPIISPPDSS
601 YLSGANLNLS CHSASNPSPQ YSWRINGIPQ QHTQVLFIAK ITPNNNGTYA
651 CFVSNLATGR NNSIVKSITV SASGTSPGLS AGATVGIMIG VLVGVALI
```

N101, N112, N149	
N179, N194	
N201, N205, N243	
N253, N271, N285, N289	
N306, N326, N347	
N356, N371,	
N428	
N462, N476	
N504, N525, N549	
N556, N576,	
N608, N646	
N661	

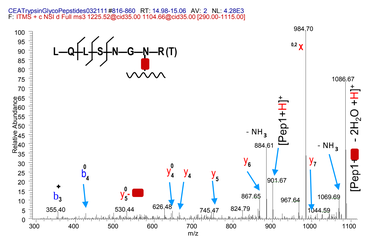
- 28 potential glycosylation sites are present in CEACAM5 protein (based on NXS/T Sequon).
- Deglycosylation by treating with PNGase F results in diagnostic deamidation of Asparagine (N) residue to Aspartic (D) acid.
- Deamidation of Asn (N) can be readily configured in a database search engine and its presence combined with the motif NXT or NXS can be used as a tool for determining N-linked glycosylation sites.
- A trypsin alone digest is performed to identify background deamidation.
- By this method, nine and/or eleven sites (N194/N549, N201, N205, N371, N476, N201/N556, N576, N646, and N661) were found to be N-glycosylated.

4. Glycan Composition Determination (N194/N549)

Annotated MS2 and MS3 [MS2 of *m/z* = 1104.7 (+1)] Spectra of *m/z* = 1225.52 (+2) Th. containing glycosylated tryptic peptide LQLSNGNR, Peptide1.



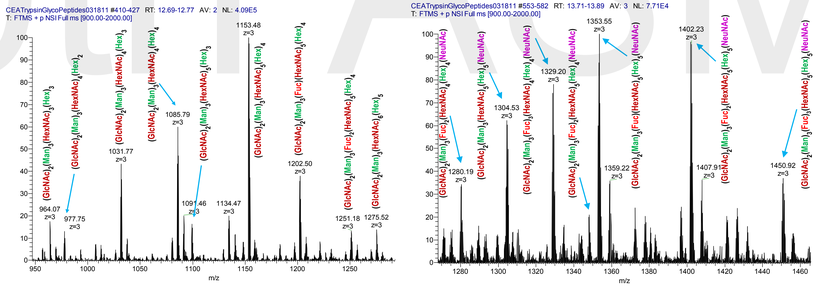
Peptide1 - $(GlcNAc)_2(Man)_3(Hex)(HexNAc)(NeuNAc)$



The validation of amino acid sequence of the Peptide1.

- MS^2 spectra of Glyco-peptides contain a wealth of information regarding the composition of glycan modification.
- Data are interpreted manually and the *m/z* ratio of the bare or minimally modified peptide is determined.
- MS^3 is performed on the minimally modified peptide to validate primary AA sequence and confirm the N-linked glycosylation site.

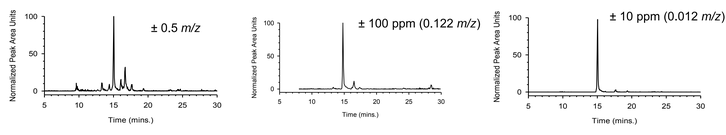
5. Heterogeneity of Glycosylation at N194/N549



- Certain other extended glycosylation forms attached to Peptide1 (N194 and/or N549): The figure on left shows glycans without NeuNAc;. The figure on right shows glycans with one or more Fucose (Fuc). The glycans containing NeuNAc but no Fucose were also observed (data not shown).
- The carbohydrate compositions indicated above were manually validated by inspection of MS^2 data.

6. HRAM Quantitation

Extracted Ion Chromatogram (XIC) of *m/z* 1225.5211



- Peak area values for different glycosylation forms of the same peptide (eg. Peptide1) can be employed for relative quantitation purposes as glycan contribution to the MS response is very minimal.
- High Resolution Accurate Mass data enables sensitive and specific quantitation of individual glycosylation forms in complex backgrounds.

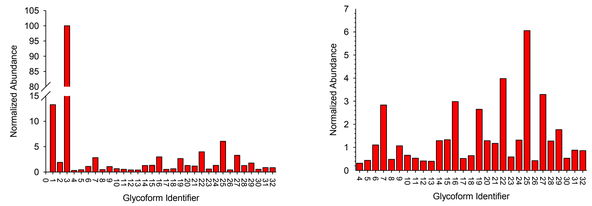
Reprints available from info@msbioworks.com

7. Glycosylation forms identified & validated: Peptide1

Glycosylation form Identifier	Experimental *m/z* (z)	Composition of the Glycans for Peptide **LQLSNGNR**	Theoretical m/z (z)	Error (ppm)
1	978.4319 (+2)	$(GlcNAc)_2(Man)_3(Hex)$	978.4308 (+2)	1.09
2	998.9454 (+2)	$(GlcNAc)_2(Man)_3(HexNAc)$	998.9441 (+2)	1.27
3	1059.4591 (+2)	$(GlcNAc)_2(Man)_3(Hex)_2$	1059.4572 (+2)	1.76
4	1078.4303 (+2)	$(GlcNAc)_2(Man)_3(Hex)(HexNAc)$	1079.9705 (+2)	0.71
5	1160.9979 (+2)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)$	1160.9969 (+2)	0.83
6	1181.5108 (+2)	$(GlcNAc)_2(Man)_3(Hex)(HexNAc)_2$	1181.5102 (+2)	0.48
7	1225.5211 (+2) & 817.3479(+3)	$(GlcNAc)_2(Man)_3(Hex)(HexNAc)(NeuNAc)$	1225.5182 (+2)	2.34
8	1242.0250 (+2)	$(GlcNAc)_2(Man)_3(Hex)_3(HexNAc)$	1242.0233 (+2)	1.34
9	1262.5378 (+2)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)_2$	1262.5366 (+2)	0.93
10	1306.5474 (+2) & 871.3656 (+3)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)(NeuNAc)$	1306.5446 (+2)	2.12
11	1327.0606 (+2) & 885.0410 (+3)	$(GlcNAc)_2(Man)_3(Hex)(HexNAc)_2(NeuNAc)$	1327.0579 (+2)	2.01
12	1387.5736 (+2) & 925.3848 (+3)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)(NeuNAc)$	1387.5710 (+2)	1.85
13	1408.0877 (+2) & 939.0605 (+3)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)_2(NeuNAc)$	1408.0843 (+2)	2.39
14	952.7355 (+3)	$(GlcNAc)_2(Man)_3(Hex)_2(HexNAc)_3(NeuNAc)$	952.7342 (+3)	1.37
15	987.7462 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_2(Fuc)(NeuNAc)$	987.7446 (+3)	1.59
16	1006.7537 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_3(NeuNAc)$	1006.7518 (+3)	1.89
17	1020.4288 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)(NeuNAc)$	1020.4273 (+3)	1.44
18	1055.4401 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_2(Fuc)(NeuNAc)$	1055.4378 (+3)	2.21
19	1060.7713 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_4(NeuNAc)$	1060.7694 (+3)	1.79
20	1074.4469 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_2(NeuNAc)$	1074.4449 (+3)	1.83
21	1109.4575 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_2(Fuc)(NeuNAc)$	1109.4554 (+3)	1.93
22	1128.4650 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_3(NeuNAc)$	1128.4625 (+3)	2.19
23	1158.1434 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_3(Fuc)_2(NeuNAc)$	1158.1413 (+3)	1.79
24	1177.1517 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_3(Hex)_3(Fuc)(NeuNAc)$	1177.1485 (+3)	2.72
25	1182.4828 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_4(NeuNAc)$	1182.4801 (+3)	2.26
26	1225.8369 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_3(Fuc)_2(NeuNAc)$	1225.8345 (+3)	1.99
27	1231.1685 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_4(Fuc)(NeuNAc)$	1231.1661 (+3)	1.95
28	1250.1751 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_5(Hex)_4(NeuNAc)$	1250.1733 (+3)	1.47
29	1279.8515 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_4(Fuc)_2(NeuNAc)$	1279.8521 (+3)	0.44
30	1298.8609 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_5(Hex)_4(Fuc)(NeuNAc)$	1298.8592 (+3)	1.29
31	1304.1937 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_5(Hex)_5(NeuNAc)$	1304.1909 (+3)	2.18
32	1328.5360 (+3)	$(GlcNAc)_2(Man)_3(HexNAc)_4(Hex)_4(Fuc)_3(NeuNAc)$	1328.5380 (+3)	1.53

- Fifty (50) Multiple glycoforms of the LQLSNGNR peptide are identified from a single purified protein sample.
- High resolution accurate mass data are essential to support the identification and characterization process

8. Glycosylation Forms Quantification



- These data show the relative abundance of the glycosylation forms (1-32) identified in Section 7.0.
- Based on these data the $(GlcNAc)_2(Man)_3(Hex)_2$) glycosylation form of LQLSNGNR is the most abundant.
- Under close analysis, there is clear variation in relative abundance data that is a function of the carbohydrate composition of the glycans.

9. Summary

- This presentation summarizes a workflow for the identification, characterization, and quantification of protein glycosylation forms.
- The combination of immuno-purification, a targeted multiple enzyme approach and high resolution accurate mass data enables the accurate identification of N-glycosylation sites and their glycosylation forms
- Glycosylation screening with PNGaseF identified nine sites as N-linked glycosylated.
- Greater degree of heterogeneity of glycosylation observed for N194/N549 (50 forms) and N201/N556 (43 forms), out of five/seven sites characterized so far. Other sites include N476 (12 forms), N371 (17 forms) and N576 (13 forms).
- The glycans which are sialylated, contain only one NeuNAc (at least in those validated so far).
- These data represent only a small portion of the quantitative data that has been derived for the different glycosylation forms of CEA and have value not only for understanding glycan biosynthesis but also for understanding the value of protein micro-heterogeneity in evaluating the disease state of CRC patients.