RAPID COMMUNICATIONS IN MASS SPECTROMETRY
Rapid Commun. Mass Spectrom. 2005; **19**: 3002–3006
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/rcm.2162

Simple modification of a protein database for mass spectral identification of N-linked glycopeptides

James A. Atwood III[1]*, Satya S. Sahoo[1,4], Gerardo Alvarez-Manilla[1], Daniel B. Weatherly[2], Kumar Kolli[3], Ron Orlando[1] and William S. York[1]

[1]Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Road, Athens, GA 30602-4712, USA
[2]Center for Tropical and Emerging Global Disease, University of Georgia, 629 Biological Sciences, Athens, GA 30602-2606, USA
[3]Windber Research Institute, 600 Somerset Ave, Windber, PA 15963, USA
[4]Large Scale Distribution Information Systems Lab., Department of Computer Science, University of Georgia, Athens, GA 30602, USA

Received 5 June 2005; Revised 27 August 2005; Accepted 27 August 2005

We describe an algorithm which modifies a protein database such that during a database search deamidation is limited to asparagines strictly contained within the N-glycosylation consensus sequence. The modified database was evaluated using a dataset created from the shotgun proteomic analysis of N-linked glycopeptides from human blood serum. We demonstrate that the application of the modified database eliminates incorrect glycopeptide assignments, reduces the peptide false-discovery rate, and eliminates the need for manual validation of glycopeptide identifications.

Asparagine-linked protein glycosylation (N-linked) is a common post-translational modification which is important in many cellular processes, including protein targeting, folding, and stability.[1] Recent studies have also demonstrated that N-linked glycosylation is significant as both an indicator and effector of disease pathogenesis.[2–4] Thus, a number of reports have emerged which detail the development of methodologies for the high-throughput identification of N-linked glycopeptides from complex mixtures such as blood serum and whole cell lysates.[5–10] However, the complexity of such protein mixtures represents a unique challenge in regards to N-linked glycopeptide analysis. Being that the proportion of N-linked glycopeptides is relatively small with respect to the total peptide population, non-glycopeptides must be depleted from the sample in order to detect the N-linked glycopeptides. Such N-linked glycopeptide enrichment has been performed via lectin affinity chromatography, hydrophilic affinity separations, and the coupling of carbohydrate-containing peptides to hydrazide resins.[5–10]

Following glycopeptide enrichment, glycopeptide analysis follows a general schema. First the carbohydrate chains are cleaved from the peptides through treatment with peptide N-glycosidase F (PNGase F). Removal of the carbohydrate moieties from the peptide backbone serves two purposes. First, the sample complexity is further reduced by removal of carbohydrate heterogeneity, and, second, deglycosylation with PNGase F induces a mass tag at the site of glycosylation by converting the previously carbohydrate-linked asparagine into an aspartic acid, a monoisotopic mass shift of 0.9840 Da. The deglycosylated peptides are then analyzed by tandem mass spectrometry (MS/MS), and software algorithms are used to correlate the experimental fragmentation spectra with theoretical tandem mass spectra generated from peptides in a protein database.[11,12] The conversion of asparagine into aspartic acid by PNGase F-catalyzed deglycosylation is accounted for by allowing asparagine deamidation as a variable modification in the database search. However, by not constraining deamidation to asparagines strictly within the N-glycosylation consensus sequence N-X_{aa}-S/T (X_{aa} is any amino acid other than proline), peptides which do not contain the consensus sequence can be identified as being formerly glycosylated. Therefore, in all previous high-throughput analyses of N-linked glycopeptides, the glycopeptide identifications made by the database search software have been manually validated to assure that they contain a deamidated asparagine residue within the N-glycosylation consensus sequence.[5–10] While this approach will result in a more accurate dataset, manual validation is both time-consuming and not amenable to high-throughput analysis.

Our work is currently focused on the development of methodologies for the high-throughput analysis of N-linked glycopeptides. Towards this goal we have designed and implemented a simple program to modify a protein database such that during a database search deamidation is limited to asparagines strictly contained within the N-glycosylation consensus sequence. The program, which is written in Java 1.5.0, iteratively searches through the amino acid sequence of each protein in a fasta database and locates the consensus sequence for N-glycosylation, N-X_{aa}-S/T represented by the regular expression {N[^P][ST]} and replaces it with an alternate representation {J[^P][ST]} which represents the

*Correspondence to: J. A. Atwood III, Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Road, Athens, GA 30602-4712, USA.
E-mail: jatwood@chem.uga.edu
Contract/grant sponsor: NIH Integrated Technology Resource for Biomedical Glycomics; contract/grant number: P41RR018502.

deglycosylated consensus sequence D-X_{aa}-S/T. Prior to database searching, the amino acid 'J' is defined as having a monoisotopic and average mass of 115.02694 and 115.0886, respectively. Thus, when the modified protein database is enzymatically digested *in silico*, this leads to the formation of theoretical peptides that correspond precisely in structure and mass to those generated by PNGase F-catalyzed deglycosylation. The database search can then be performed using no modifications and only peptides containing asparagine residues within the N-glycosylation consensus sequence will be considered as modified by deamidation via the presence of the residue 'J'.

EXPERIMENTAL

Preparation of serum peptides

For this study a test dataset of both N-linked glycopeptides and non-glycopeptides was generated from depleted human serum (Sigma) as previously described.[13] Briefly, the human serum (1 mL) was depleted of albumin and IgG by first passing it over a blue sepharose resin (GE-Amersham Biosciences, Piscataway, NJ, USA) then over protein G agarose (GE-Amersham Biosciences). The depleted serum was lyophilized then dissolved in a solution of 0.2% RapiGest (Waters, Milford, MA, USA) in 50 mM NH_4HCO_3. The proteins were then reduced (25 mM dithiothreitol (DTT), pH 8.5) for 30 min at 45°C followed by carboxyamidomethylation with iodoacetamide (90 mM) for 1 h at room temperature in the dark. Samples were then digested overnight at 37°C with 33 μg TPCK-treated trypsin (Pierce, Rockford, IL, USA). The surfactant was removed by acidification with HCl followed by centrifugation at 14 000 *g*.

Peptide separation and LC/MS/MS analysis

The peptide mixture was directly desalted and separated into two fractions based on peptide size over a 3.2 × 30 cm Superdex peptide column (GE-Amersham Biosciences), eluted isocratically with an aqueous solution that contained 0.1% trifluoroacetic acid (TFA) at a flow rate of 0.1 mL/min. The peptides in each fraction were deglycosylated overnight by incubation at 37°C with 0.2 units of PNGase F (Sigma). The fractions were then dried under vacuum, resuspended in 100 uL of 0.1% TFA, injected onto a 1 × 150 mm Zorbax C_{18} column (Agilent, Palo Alto, CA, USA) at a flow rate of 50 μL/min, and separated into ten additional fractions by reverse-phase chromatography, as previously described.[14] Each fraction was dried under vacuum and resuspended in 40 μL of 0.1% formic acid for analysis by liquid chromatography/tandem mass spectrometry (LC/MS/MS). A volume of 5 μL of each fraction was analyzed independently using a Waters CapLC (Milford, MA, USA) interfaced directly to a QTOF-2 tandem mass spectrometer (Micromass, UK) as described previously.[14] Finally, the raw mass spectra were converted into peak-list format and combined prior to database searching.

Protein sequence databases and database searching

Four sequence databases were constructed for this analysis. First, a representative database (normal) was created which consisted of 27 960 *Homo sapiens* protein sequences from NCBI.[15] Using the normal database, a second database (modified) was created using a Java program which iteratively searches through the amino acid sequence of each protein in the fasta database and locates the consensus sequence for N-glycosylation, N-X_{aa}-S/T represented by the regular expression {N[^P][ST]} where X_{aa} is any amino acid other than proline. The asparagine contained within the consensus sequence is then replaced with a J using the alternate representation {J[^P][ST]} which represents the deglycosylated consensus sequence D-X_{aa}-S/T. The amino acid 'J' was then defined as having a monoisotopic and average mass of 115.02694 and 115.0886, respectively. Two decoy databases (reverse normal and reverse modified) were also created by reversing the protein sequences in both the normal and modified databases. The concatenated peak-list was then independently searched against all four databases using Mascot (version 1.9, Matrix Science, London, UK) with the following parameters: fully tryptic enzymatic cleavage, two allowed missed cleavages, peptide tolerance of 60 parts-per-million, 0.2 Da fragment ion tolerance, and a variable modification of cysteine (+57 Da). For the searches against the normal and reversed normal databases, deamidation of asparagine (+0.9840 Da) was also allowed as a variable modification.

Peptide identification and false-discovery rate calculations

Peptide matches above discrete Mascot ion scores were extracted from the normal, reverse normal, modified, and reverse modified database search results. Peptide redundancy was removed and the peptide false-discovery rates (PEP-FDR) were then calculated at each ion score threshold for both the normal and modified database results as previously described.[14] Peptides which matched with scores below a 5% PEP-FDR in both the normal and modified databases were then separately clustered to proteins forming a list of peptide and protein identifications resulting from each database search.[14] The distribution of the peptide and protein identifications below a 5% PEP-FDR from both the normal and modified databases are shown in Table 1.

Table 1. Distribution of peptide and protein assignments below a 5% peptide false-discovery rate from the normal and modified database search results. False glycopeptide and glycoprotein identifications do not occur in the modified database search

	Normal database	Modified database (J-X-S/T)
Unique peptides	206	199
Unique glycopeptides	26	20
Unique glycopeptides containing motif (NXS/T)	20	20
Unique proteins	34	29
Unique glycoproteins	15	11
Unique glycoproteins containing motif (NXST)	11	11

RESULTS AND DISCUSSION

Peptide false-discovery rate

As the effective database size increases through either an increase in the number of allowed modifications or an increase in the allowed mass tolerance, the score distribution will shift to higher scores when searching a decoy database.[16–18] The implications of this are two-fold. First, the probability of obtaining an incorrect peptide identification at a significant score will increase, and second the score threshold required to maintain a specified false-discovery rate (FDR) will also increase. This would suggest that by limiting the number of allowed modifications in a database search that one could employ lower score thresholds for protein or peptide identification, and the frequency of significant random matches would also decrease. Thus, when performing a database search for formerly N-linked glycopeptides, the frequency of random glycopeptide assignments should decrease by searching against the modified database described above. Figure 1 shows the distribution of peptide FDRs as a function of Mascot ion scores for both the normal and modified database search results. Also included in Fig. 1 is the frequency of incorrectly assigned glycopeptides at or above discrete Mascot ion scores from the normal database search. Incorrectly identified glycopeptides were defined as peptides which did not contain the consensus sequence for N-linked glycosylation but were reported to have a deamidated asparagine. From Fig. 1 it is evident that the frequency of false discoveries at each score threshold is slightly decreased when searching the modified database versus the normal database. A PEP-FDR is achieved in the modified database by selection of peptides exceeding ion scores of 33. To achieve the same error rate in the normal database an ion score threshold of 36 would be necessary. We interpret this decrease in the PEP-FDR to be a result of the restriction of possible asparagine modifications to asparagines contained within the N-linked glycosylation consensus sequence. While the PEP-FDR is increased in the normal database search, this increase is not due to the presence of incorrectly assigned glycopeptides. Rather, in the normal database search, in which deamidation is considered on all peptides containing an asparagine, the effective database size is larger than in the modified database and thus the likelihood of obtaining a false peptide assignment is increased.

Incorrect glycopeptide assignments–deamidation

By considering modifications to all possible asparagines, the normal database search allows for the identification of glycopeptides which do not contain the consensus sequence for N-linked glycosylation (Table 1, Fig. 1). For our analysis this resulted in the incorrect identification of six peptides as being deglycosylated below a PEP-FDR of 5% (Table 1, Fig. 1). Following manual interpretation of each incorrectly identified glycopeptide, it was determined that these assignments resulted primarily from two phenomena. First, peptides containing an asparagine which is truly deamidated but not present within the N-glycosylation consensus sequence will be identified as glycopeptides. Figure 2 is a fragmentation spectrum which resulted from the collision-induced dissociation (CID) of a doubly charged precursor ion at *m/z* 518.26 (Fig. 2, insert). The fragmentation spectrum matched with a high

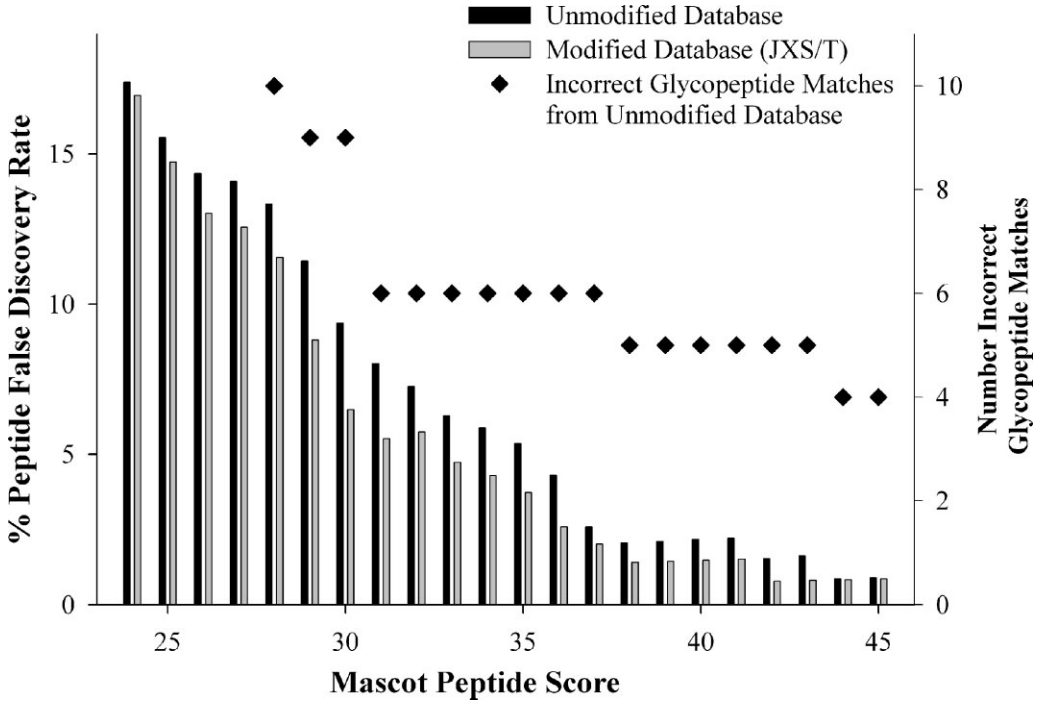

Figure 1. Comparison of the peptide false-discovery rates between the normal and modified database search results as a function of Mascot ion score threshold. By searching the modified database the effective database size is smaller, thus the peptide false-discovery rate is decreased at each ion score threshold. Conversely, in the unmodified database search, an increase in the proportion of random matches is observed and the frequency of incorrect (not containing N-X_{aa}-S/T) N-linked glycopeptide identifications increases with decreasing ion score thresholds.

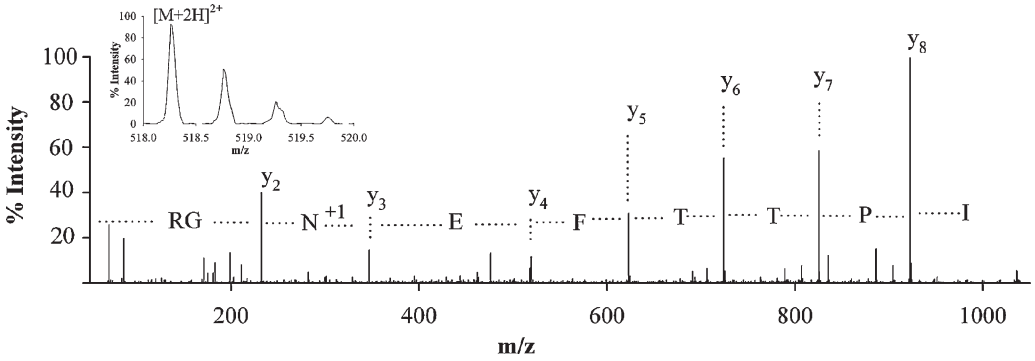

Figure 2. Positive ion CID spectrum (precursor ion $[M+2H]^{2+}$, *m/z* 518.26, in insert) from the deamidated peptide IPTTFEN^{+1}RG. N^{+1} indicates the presence of a deamidated asparagine.

confidence ion score of 63 to the deamidated peptide IPTTFEN^{+1}GR in the normal database search in which N^{+1} indicates the site of asparagine deamidation. Since deamidation of asparagine residues is known to occur both naturally and as a chemical artifact, it is not surprising that the database search allowing for deamidation of all asparagines resulted in a mixture of both true glycopeptides and incorrect glycopeptide identifications. By searching the modified database, the deamidated non-glycopeptides are easily filtered out, as observed by the absence of these identifications in the modified database search results (Table 1). To remove the false positive identifications of N-linked glycopeptides due to deamidation of asparagines, recent reports have described PNGase F deglycosylation in the presence of $H_2{}^{18}O$. Through this method, a stable ^{18}O label is introduced into the deglycosylated asparagine and a mass shift of 3 Da is observed on the deglycosylated asparagines.[19] Database searching against the modified 'J' database is also applicable in this case. By searching using a variable modification to the residue 'J' of 2 Da, the ^{18}O label is strictly assigned to asparagines found only in the N-glycosylation consensus sequence. Similarly, other modifications intended to label former sites of glycosylation can be accommodated simply by modifying the mass of the 'J' residue.

Incorrect glycopeptide assignments–precursor selection

The second source of incorrect glycopeptide identifications results from the selection of precursor ions by the mass spectrometer. In general, when data-dependent acquisition is performed, a given number of the most intense precursor ions from the survey scan are selected for subsequent MS/MS analysis. However, for ions in which the ^{13}C isotope is more intense than the monoisotopic peak, the ^{13}C isotope will occasionally be selected for MS/MS. In such cases the peptide precursor mass will be reported as 1 Da higher than the true precursor mass, and subsequently the peptide, if it contains an asparagine, could be identified as deamidated. An example of this is shown in Fig. 3. The insert in Fig. 3 indicates that the ^{13}C isotope of a doubly charged precursor ion with a *m/z* value of 1287.64 was fragmented, producing an MS/MS spectrum which matched with a high confidence ion score of 75 in the normal database to the deamidated peptide sequence TLN^{+1}QPDSQLQLTTGNGLFLSEGLK. In this dataset, the

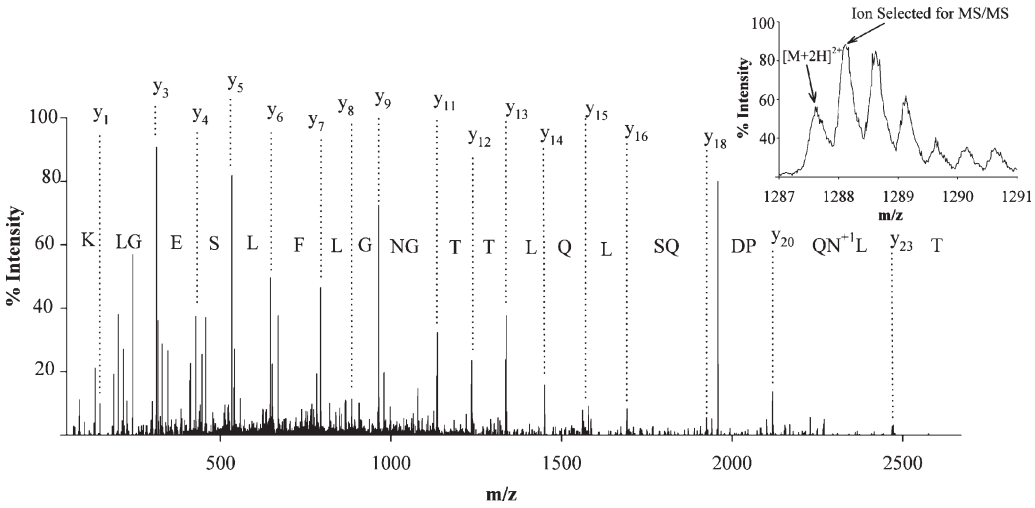

Figure 3. Positive ion CID spectrum from the incorrectly identified deamidated peptide TLN^{+1}QPDSQLQLTTGNGLFLSEGLK in which N^{+1} represents the site of deamidation. The insert shows that the ^{13}C isotope (*m/z* 1288.14) was selected for fragmentation rather than the true precursor ion ($[M+2H]^{2+}$, *m/z* 1287.64).

incorrect identification of glycopeptides via errors in precursor mass selection occurred with peptides exceeding 2000 Da which contained an asparagine at a position within the peptide backbone such that a sufficient number of fragment ions could be identified within the specified fragment ion mass tolerance by the database searching software. For example, the peptide displayed in Fig. 3 matched with high confidence because N3 was treated as deamidated. Therefore, the ion series y_1–y_{20} would not exhibit a 1 Da mass shift. However, if N15 would have been treated as deamidated, the y_{11}–y_{23} would not have matched due to an expected 1 Da mass shift, and the ion score would have been less significant. While incorrect glycopeptide assignments arise from this issue in the normal database search they were not found to occur in the modified database search (Table 1).

Glycoprotein identification

A significant difficulty associated with the analysis of glycopeptide-enriched samples is that a relatively small number of peptides (often only one) are used to make a protein identification. This is important because, as the number of incorrect glycopeptide assignments increases, the potential for erroneous identification of glycoproteins also increases. As seen in Table 1, database searching against a normal database resulted in the incorrect identification of six glycopeptides and four glycoproteins. In comparison, when the database search was performed against the modified database, these incorrect identifications were filtered out, leaving only identifications of glycopeptides and glycoproteins which contained the N-linked glycosylation consensus sequence. One important caveat is that while searching the modified database will filter incorrectly identified glycopeptides it will also prevent the identification of peptides such as ALGISPFHEHAEVVFTA***NDS***GPR which are not deamidated but do contain the N-linked glycosylation sequence. While this occurred only once in this test dataset the missed peptide identification resulted in the loss of one protein identification (Table 1). However, this is not a major concern when the purpose of the study is strictly the identification of N-linked glycopeptides.

CONCLUSIONS

In conclusion, this modification to a protein database provides a simple method to filter out theoretical peptides that could not arise by PNGase F deglycosylation while removing the tedious process of manual validation. By defining the mass of the residue 'J' as the mass of a modified asparagine the modification due to deglycosylation is limited to only asparagines contained within the N-glycosylation consensus sequence. We have demonstrated that this procedure eliminates incorrect glycopeptide assignments while reducing the peptide false-discovery rate when compared to unmodified database searches. The software used to perform the N to J and J to N conversion of a fasta protein database is publicly available.[20]

Acknowledgements

This work was supported by the NIH Integrated Technology Resource for Biomedical Glycomics (Grant P41RR018502).

REFERENCES

1. Kukuruzinska M. *Crit. Rev. Oral Biol. Med*. 1998; **9**: 415.
2. Fernandes B, Sagman U, Auger M, Demetrio M, Dennis J. *Cancer Res*. 1991; **51**: 718.
3. Etzioni A, Frydman M, Pollack S, Avidor I, Phillips M, Paulson J, Gershoni-Baruch R. *N. Engl. J. Med*. 1992; **327**: 1789.
4. Seelentag W, Li W, Schmitz S, Metzger U, Aeberhard P, Heitz P, Roth J. *Cancer Res*. 1998; **58**: 5559.
5. Xiong L, Regnier F. *J. Chromatogr. B* 2002; **782**: 405.
6. Xiong L, Andrews D, Regnier F. *J. Proteome Res*. 2003; **2**: 618.
7. Zhang H, Xiao-jun L, Martin D, Aebersold R. *Nat. Biotechnol*. 2003; **21**: 660.
8. Wada Y, Tajiri M, Yoshida S. *Anal. Chem*. 2004; **76**: 6560.
9. Brunkenborg J, Pilch B, Podtelejnikov A, Wisniewski J. *Proteomics* 2004; **4**: 454.
10. Zhang H, Yi E, Xiao-jun L, Mallick P, Kelly-Spratt K, Masselon C, Camp D, Smith R, Kemp C, Aebersold R. *Mol. Cell. Prot*. 2005; **4**: 144.
11. Perkins D, Pappin D, Creasy D, Cottrell S. *Electrophoresis* 1999; **20**: 3551.
12. Eng J, McCormack A, Yates J. *J. Am. Soc. Mass. Spectrom*. 1994; **5**: 976.
13. Alvarez-Manilla G, Atwood J, Guo Y, Warren N, Pierce M, Orlando R. *J. Protein Res*. 2005; submitted.
14. Weatherly B, Atwood J, Minning T, Cavola C, Tarleton R, Orlando R. *Mol. Cell. Prot*. 2005; **4**: 762.
15. www.ncbi.nih.gov.
16. Resing K, Meyer-Arendt K, Mendoza A, Aveline-Wolf L, Jonscher K, Pierce K, Old W, Cheung H, Russell S, Wattaw J, Goehle G, Knight R, Ahn N. *Anal. Chem*. 2004; **76**: 3556.
17. Cargile B, Bundy J, Stephenson J. *J. Proteome Res*. 2004; **3**: 1082.
18. Olsen J, Ong S, Mann M. *Mol. Cell. Prot*. 2004; **3**: 608.
19. Kristiansen T, Bunkenborg J, Gronborg M, Molina H, Thuluvath P, Argani P, Goggins M, Maitra A, Pandey A. *Mol. Cell. Prot*. 2004; **3**: 715.
20. Available: http://128.192.9.86/stargate/ModifyDB.zip.