ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Statistically appraising process quality of affinity isolation experiments

Julia L. Sharp [a,*], John J. Borkowski [b], Denise Schmoyer [c], Don S. Daly [d], Samuel Purvine [d], William R. Cannon [d], Gregory B. Hurst [c]

[a] Clemson University, 237 Barre Hall, Clemson, SC, USA

[b] Montana State University, 2-214 Wilson Hall, Bozeman, MT, USA

[c] Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN, USA

[d] Pacific Northwest National Laboratory, 900 Battelle Blvd, PO Box 999, Richland, WA, USA

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

Quality affinity isolation experiments are necessary to identify valid protein–protein interactions. Biological error, processing error, and random variability can reduce the quality of an experiment, and thus hinder the identification of protein interaction pairs. Appraising affinity isolation assay quality is essential to inferring protein associations. An important step of the assay is the mass spectrometric identification of proteins. To evaluate this step, a known mixture of proteins is processed through a mass spectrometer as a quality control mixture. If the mass spectrometer yields unexpected results, the process is currently qualitatively evaluated, tuned, and reset. Statistical quality control (SQC) procedures, including the use of cumulative sum, the individual measurement, and moving range charts are implemented to analyze the stability of the mass spectrometric analysis. The SQC measures presented can assist in establishing preliminary control limits to identify an out-of-control process and investigate assignable causes for shifts in the process mean in real time.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

One approach for identification of valid protein–protein interactions necessitates the use of high-quality affinity isolation experiments. Currently, the majority of effort has focused on the production of large quantities of data from these experiments. If, however, contamination is introduced into the experimental process, then the data are unreliable and possibly misleading. Specifically, inferred interactions between protein pairs may not be statistically valid.

In this paper, a statistical quality control (SQC) methodology adapted for use in the mass spectrometric identification of proteins will be presented. This step is an important component of one common type of affinity isolation experiments (Berggard et al., 2007), as well as other proteomics-related measurements (Smith et al., 2007). These statistical methods are well-suited for this type of experimental data, and will assist in the establishment of preliminary control limits that are used to signal an out-of-control process. Once these limits are established, assignable causes for shifts in the process mean can be found in real time.

2. Background

For the experimental process examined in this paper, one goal is to achieve a state of statistical process control. Researchers aim to manage and establish control of the process by analyzing a "known" protein sample with the liquid

* Corresponding address: Department of Applied Economics and Statistics, Clemson University, 237 Barre Hall, Clemson, SC, 29634, USA. Tel.: +1 864 656 3252; fax: +1 864 656 5776.

E-mail address: jsharp@clemson.edu (J.L. Sharp).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.011

Please cite this article in press as: Sharp, J.L., et al., Statistically appraising process quality of affinity isolation experiments. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.011

Table 1
Peptide and protein labels from the standard mixture of proteins

Peptide/Protein	
Label	Name
ADH2_YEAST	Yeast alcohol dehydrogenase II
ADH1_YEAST	Yeast alcohol dehydrogenase I
HBA_BOVIN	Bovine hemoglobin α
HBB_BOVIN	Bovine hemoglobin β
ALBU_BOVIN	Bovine serum albumin
MYG_HORSE	Horse myoglobin
LYC_CHICK	Chicken egg lysozyme C
CAH2_BOVIN	Bovine Carbonic anhydrase 2

chromatograph/mass spectrometer (LC-MS) system. This known sample is a standard mixture containing fixed quantities of certain peptides run through the mass spectrometer along with experimental runs. Specifically, data were collected from LC-MS analyses of this standard mixture as part of a "pipeline" for measurement of protein–protein interactions in *R. palustris* at Oak Ridge National Laboratory from July, 2004 to May, 2007. Each sample of the quality control standard is a mixture containing the peptides resulting from tryptic digestion of eight proteins which are listed (with corresponding abbreviated labels) in Table 1 (Strader et al., 2006). These standard samples are to be processed through the mass spectrometer both before and after a batch of experimental samples through LC-MS system. Because a standard sample is formulated as a known mixture, the results from the mass spectrometer analysis should be similar from one standard sample to another. If, however, the mass spectrometer yields results that are inconsistent or unusual with respect to the components of the standard mixture, then the process is evaluated, tuned, and reset if necessary.

Several aspects of the data from LC-MS analyses of the standard can be evaluated to determine whether the operating performance of the LC-MS system is acceptable performance. Reproducibility of LC retention times of peaks for a small set of selected peptides provides evidence that the liquid chromatography components of the system are functioning properly. Shifted retention times or missing peaks can indicate hardware problems with the LC system (worn parts, leaks, etc.) as well as incorrectly prepared mobile phase solutions. Peaks with widths that are wider than expected can indicate a need for replacement of the LC column. Identification of fewer peptides than expected can indicate poor performance of the mass spectrometer itself, suggesting the need for calibration, tuning, cleaning, replacement of detector elements, or other similar maintenance procedures. This paper focuses on monitoring the number of identified peptides in a standard mixture as an indicator of the performance of the LC-MS system.

Statistical quality control (SQC) procedures applied to the affinity isolation experiment data will establish a quantitatively rigorous approach to the identification of an out-of-control process. The implementation of the procedures will give researchers an objective methodology for monitoring the process. The SQC procedures presented are also capable of detecting small shifts in the process, a practical consideration for analyzing data from LC-MS analyses.

3. Methods

For any quantitative process response of interest, there are two sources of variation to be monitored. First, there is the random variability inherent to the process which is represented by σ^2, the process variance. Second, there is variability in the process mean μ which can result from short-term or instantaneous changes in the mean, as well as from long-term trends or gradual changes in the mean. Ideally, σ^2 will be acceptably small and μ will be equal to an aim value, denoted μ_0.

To analyze the stability of the affinity isolation experiment process in terms of these two sources of variation, SQC charting techniques will be used. Specifically, the cumulative sum (cusum) chart, the individual measurements (I) chart, and the moving range (MR) chart will be implemented. A cusum chart is a plot of the cumulative deviations of a sequence of observations from an aim value, usually the process mean. The individual measurements or I-chart is a sequence plot of the measurements against the order of data acquisition. Both the cusum chart and I-chart are used to monitor the process mean μ. The cusum is designed to detect relatively small shifts in μ faster than the I-chart while the I-chart will detect large shifts in μ faster than the cusum chart. The moving range (MR) chart is used to monitor σ^2, the random variability inherent to the process (Montgomery, 2005).

If a cusum or I-chart indicates an "out-of-control" signal, then we conclude a shift in μ has occurred. If a MR-chart indicates an "out-of-control" signal, then we conclude a shift in σ^2 has occurred. Until an out-of-control signal occurs, the process is considered to be in a state of statistical control. For the affinity isolation experiments, cusum charts and I-charts will be used to monitor the mean of the peptide counts collected from the processed standard protein mixtures. When the process is "in-control", the differences between the individual observations and some aim peptide count will be distributed with expectation zero (Montgomery, 2005). Thus, when the process is in-control, the points on the cusum chart should be close to zero and the points on the I-chart should be randomly distributed about the aim. Although beyond the scope of the present paper, additional aspects of the data measurement from LC-MS analyses (retention times, for example) could be included in future implementations of the statistical analysis of performance.

Please cite this article in press as: Sharp, J.L., et al., Statistically appraising process quality of affinity isolation experiments. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.011

If the process variance σ^2 is unknown or cannot be specified by the researcher for any experimental result of interest, then it must be estimated from the data. The sample variance s^2 is not a good estimator of the σ^2 for an in-control process because it reflects not only inherent process variability, but also incorporates measurement error and long-term variability due to assignable causes. Thus, s^2, on average, over-estimates σ^2. Instead of s^2, the mean squared successive difference (mssd) estimator $\widehat{\sigma}^2$ is the most commonly-used estimator of σ^2, and is the default method used for the cusum procedure in SAS (SAS Institute Inc., 1995). The mssd estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{2(N-1)} \sum_{i=1}^{N-1} (y_{i+1} - y_i)^2, \tag{1}$$

where y_i is the ith observation and N is the number of individual observations (SAS Institute Inc., 1995). The mean squared successive difference is, in general, a good estimate of the process variability and measurement error because the effect of long-term variability due to assignable causes is minimized by using only successive observations (von Neumann et al., 1941).

Because individual measurements (peptide counts) are recorded for a given protein, the I-chart (individual measurements chart) and the MR-chart (moving range chart) are useful for displaying the two components of process variability: variability in the process mean μ and the random process variability σ^2, respectively.

A moving range is computed by finding the absolute difference between two consecutive observations. The ith moving range is given by $MR_i = |y_{i+1} - y_i|$. The estimator of the process variability in the moving range chart is $\hat{\sigma} = \frac{\overline{MR}}{d_2}$ where $\overline{MR} = \left(\sum_{i=1}^{N-1} MR_i\right)/(N-1)$ is the average of the $N-1$ consecutive moving ranges in N observations and $d_2 \approx 1.128$ is the mean of the distribution of the relative range $(y_{\max} - y_{\min})/\sigma$ under normality assumptions with a sample size of two (Montgomery, 2005).

The standard 6σ range between a lower control limit (LCL) and an upper control limit (UCL) will be adopted for both the I and MR control charts. Then LCL and UCL for the MR-chart are defined to be

$$LCL = \overline{MR}\left(1 - \frac{3d_3}{d_2}\right) \quad \text{and} \quad UCL = \overline{MR}\left(1 + \frac{3d_3}{d_2}\right) \tag{2}$$

where $d_3 \approx 0.853$ is the standard deviation of the distribution of the relative range. If the LCL for the MR chart is < 0 then it is set to 0.

For the I-chart, the LCL and UCL for individual measurements are

$$LCL = \bar{y} - 3\frac{\overline{MR}}{d_2} \quad \text{and} \quad UCL = \bar{y} + 3\frac{\overline{MR}}{d_2}. \tag{3}$$

To establish a basis from which a more refined set of aim values can be developed, this exploratory data analysis will use sample means of peptide counts as preliminary aim values. Thus, for any peptide, the average difference between individual observations and the sample mean will be zero. Therefore, we can use cusum charts to detect any unusual patterns or systematic deviations from randomness for any particular peptide count (e.g., when the process mean for a peptide count has shifted away from the preliminary aim value). In addition, once a shift in a mean has been detected, a cusum-based estimate of the current mean can be found.

For a particular sample of the standard mixture (identified by the job identification number), the recorded measurement is a count of the number of peptides identified from the mass spectrometry data. Suppose a total of N standard samples is processed. Let y_i be the total number of peptides observed for the ith job identification number. The ith cusum (S_i) is defined to be the sum of the deviations of the first i observations from the process aim value (Hawkins and Olwell, 1998; Montgomery, 2005). Initially, using the sample mean $\bar{y}$ as the preliminary aim, the ith cusum is

$$S_i = \sum_{j=1}^{i} (y_j - \bar{y}). \tag{4}$$

Let $\max(a, b)$ be the maximum of a and b. The ith cusum for detecting an upward shift (S_i^+) for the ith job identification number is defined as

$$S_i^+ = \max(0, y_i - K^+ + S_{i-1}^+) \tag{5}$$

where $K^+ = k\sigma$ is the reference value of the "upward" cusum and σ^2 is the process variance. The value of k is chosen relative to the size of the shift to be detected and S_0^+ is the head start value for the upward shift (usually set to 0). Similarly, the ith cusum for detecting a downward shift (S_i^-) for the ith job identification number is

$$S_i^- = \max(0, -K^- - y_i + S_{i-1}^-) \tag{6}$$

where $K^- = k\sigma$ is the reference value of the "downward" cusum (Lucas, 1985; Hawkins and Olwell, 1998; Montgomery, 2005). S_0^- is the head start value for the downward shift, and is usually set to 0.

The cusum parameter values K^+ and K^- are, respectively, values above and below which the counted data cusums are not designed to react. An out-of-control signal occurs on the ith sample when either S_i^+ or S_i^- reaches or exceeds the decision

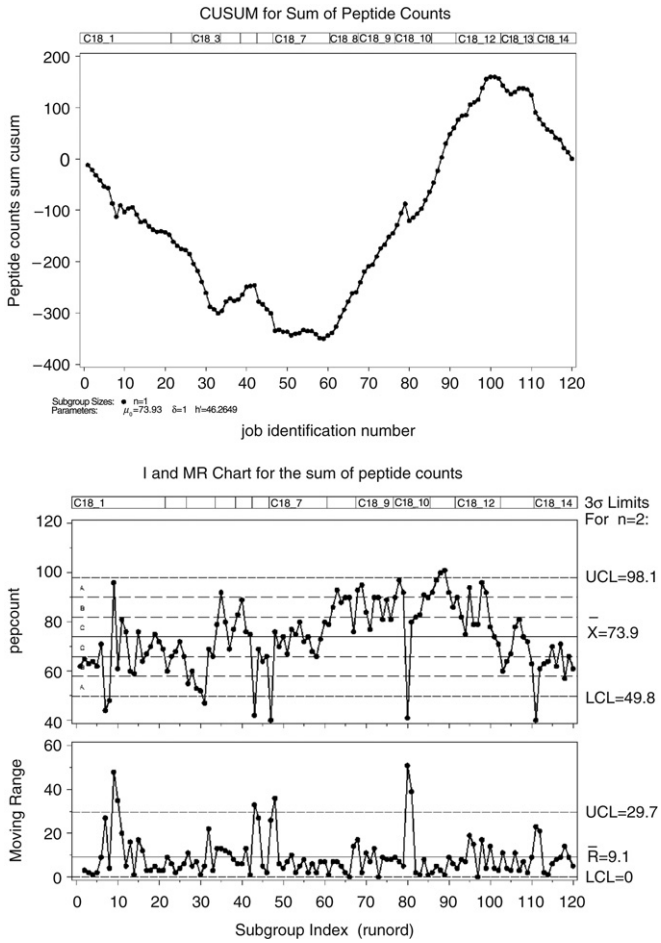

Fig. 1. Cumulative sum, individual observations, and moving range control charts for the sum of peptide counts that occur in a particular job identification number.

interval value, $H = h\sigma$. The goal is choose h and k to optimize the in-control average run length (ARL), denoted ARL_0, to be sufficiently large when the process mean is on-aim and sufficiently small when the process mean shifts $\pm\Delta$ units from the on-aim mean value for some pre-specified value Δ. Hawkins and Olwell (1998) provide guidance for the choices of h, k, and ARL. In addition, Montgomery (2005) explains that in order to detect a one standard deviation shift in the process, the use of $h = 4$ or $h = 5$ and $k = 0.5$ provides a cusum with an appropriately long ARL.

In addition to considering the cusum and I and MR charts, investigators also might want to consider 'runs rules'. These rules, or tests, detect particular non-random patterns in the data and not just signals caused by an individual outlying value. In the I chart, zones are defined to assist in the identification of these non-random patterns. Zones C, B, and A are defined, respectively, to be the areas on the control chart that are (i) within $\pm 1\sigma$ of the aim, (ii) between -2σ and -1σ or between 1σ and 2σ from the aim, and (iii) between -3σ and -2σ or between 2σ and 3σ from the aim. As an example of a runs rule, investigators might want to be warned if seven consecutive observations fall outside of zone C on the I chart. If σ is estimated, then the zones are formed using the estimate $\frac{\overline{MR}}{d_2}$.

4. Results

From July, 2004 to May, 2007, data were collected at the Oak Ridge National Laboratory from each sample of the quality control standard protein mixture (Strader et al., 2006). Each standard sample has a unique job identification number that is linked to the mass spectrometer analysis number. The job identification numbers are ordered according to their acquisition start time (i.e., the time that the sample entered the mass spectrometer for analysis). A particular job identification number has an associated list of peptides that are identified with a particular protein, and the number of peptides per protein is tabulated. Using Sequest (Eng et al., 1994; Eng and Yates, 1999), peptides are identified via automated cross-correlation-based matching of experimental tandem mass spectra and predicted tandem mass spectra of the expected peptides in the sample. Default threshold values for Sequest results were applied, with $XCorr \geq 1.8$ for singly-charged peptides, $XCorr \geq 2.5$ for doubly charged peptides, and $XCorr \geq 3.5$ for triply-charged peptides, and $DeltCN \geq 0.08$ (Tabb et al., 2002).

Please cite this article in press as: Sharp, J.L., et al., Statistically appraising process quality of affinity isolation experiments. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.011

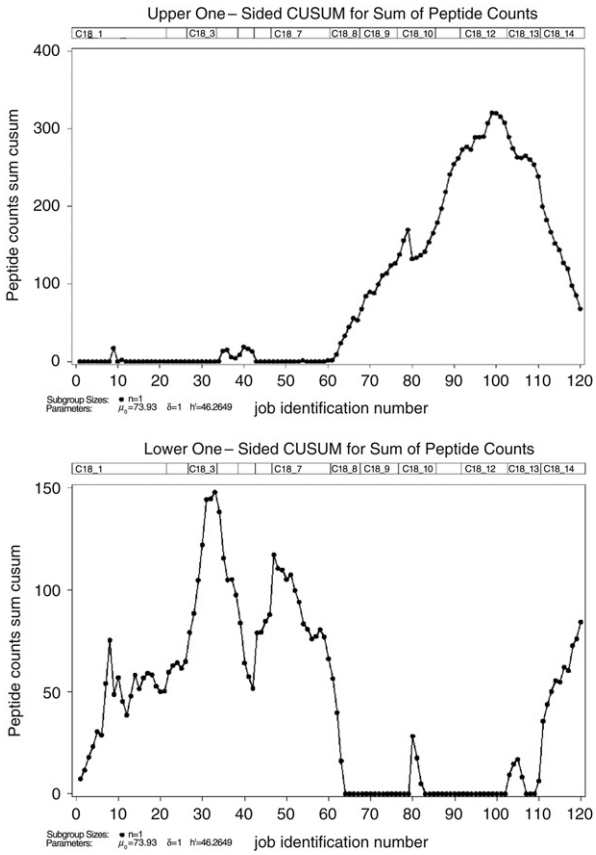

Fig. 2. Upper and lower one-sided cumulative sum control charts for the sum of peptide counts from the standard protein mixture that occur in a particular job identification number.

Although each protein can be analyzed separately, the sum of peptide counts for each job identification number is analyzed in the following example. Figs. 1 and 2 contain the S_i cusum chart, the I and MR-charts, and the upper S_i^+ and lower S_i^- one-sided cusum control charts for the sum of peptide counts across job identification numbers. In Fig. 1, the pattern for the mean for the sum of peptide counts is quite similar to the pattern that was observed for many of the individual peptides (not shown).

The sample mean (and subsequent aim value) of the 120 standard protein mixture samples from this analysis is 73.93 peptides. The mass spectrometer column used to separate the components of the digested mixture is labeled in the segmented bar above each of the plots. The cusum, upper cusum, and lower cusum charts show that the process mean is operating below the aim value for most of the first half of the standard protein mixture samples. The process appears to be operating around the aim (73.93 peptides) from about the 9th to the 12th sample and again around the 47th to the 61st sample (Fig. 1). The peptide counts associated with these latter samples were measured using the mass spectrometer column $C18_7$. Upon a change in the mass spectrometer column to $C18_8$, the mean of the process shifts above the aim value. The process operates above the aim value until columns $C18_12$ and $C18_13$ are used in the mass spectrometer and the process appears to stabilize. With a column change to column $C18_14$, the process shifts to operate below the aim value. The linear trends in the cusum plot indicate that most of the shifts in the process mean that occur in this analysis are constant shifts. Non-linear trends indicate dynamic changes have occurred in the process mean.

Out-of-control signals are observed in the I and MR charts. An out-of-control signal will be observed when the individual measurements are outside of three standard deviations from the mean. In the I-chart, job identification numbers that occur outside of three standard deviations from the mean are: 6, 7, 31, 43, 47, 80, 88, 89, and 113. Moving range values that are outside the upper control limit occur for the job identification numbers 9, 10, 43, 48, 80, and 81. While out-of-control signals are evident in the I and MR control charts, the control limits for the moving range ($LCL = 0,\ UCL = 29.7$) and for the individual measurements ($LCL = 49.8,\ UCL = 98.1$) are calculated using all of the observations. However, in the presence of outliers, it is common to revise the trial control limits by recalculating a revised mssd estimate of σ^2 after the removal of outliers.

For data from LC-MS analyses such as those described in this paper, the estimated rate of incorrect peptide identification is less than 5%. These incorrect identifications are based on Sequest searches using a "decoy" database containing both normal and amino acid sequence-reversed versions of all *R. palustris* proteins (data not shown); an identification event for

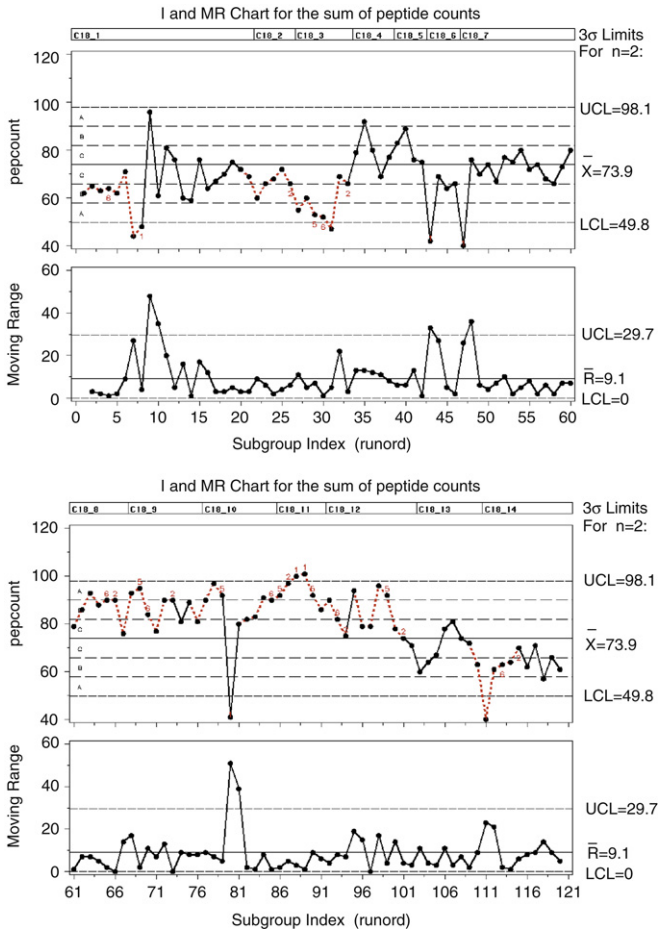

Fig. 3. Individual observations and moving range charts with special runs tests flagged for the sum of peptide counts that occur in a particular job identification number. The indices in the figure state which runs rule has been violated.

a "reverse" peptide is considered incorrect (Elias and Gygi, 2007). For the present paper, Sequest searches were performed using a protein database containing amino acid sequences for the proteins in the standard mixture, as well as for more than 4800 *R. palustris* proteins. The standard samples that are the subject of the present paper do not contain any *R. palustris* proteins. Any peptides identified in the standard protein mixture as arising from *R. palustris* proteins could be considered either incorrect identifications, or due to instrument carryover from previous samples.

In the one hundred twenty standard protein mixture samples through the mass spectrometer, only eight had no *R. palustris* identifications. The other runs had at least one *R. palustris* identification. One of the standard protein mixture samples considered to be out-of-control using the *I* and *MR* charts (job identification number 43) had no *R. palustris* identifications. Twenty-six standard protein mixture samples had more than five *R. palustris* peptides identifications, only three of which corresponded to out-of-control signals (job identification numbers 6, 88, and 89) on the *I* and *MR* charts. These results indicate that standard protein mixture samples registering out-of-control signals did not coincide with the observation of higher false-positive peptide identification rates.

Fig. 3 illustrates the use of runs rule tests for the sum of peptide counts that occur in a particular job identification number. The violation of a runs rule is flagged by an index in the figure. Specifically, in the *I*-chart, if one peptide count sum occurs beyond Zone A then a '1' occurs in the figure. A '2' occurs if seven consecutive points occur in Zone C or beyond on one side of the aim value line. A '5' occurs if two out of three consecutive points are in Zone A or beyond. A '6' occurs when four out of five sequential points are in Zone B or beyond. Additional runs rules are available in SAS (SAS Institute Inc., 2000–2004, 1995).

5. Conclusion

Statistical quality control procedures applied to affinity isolation experiment data assist the researchers in establishing a quantitatively rigorous approach to identify an out-of-control process. The cusum, *I*, and *MR* control charts presented here can assist researchers to monitor the stability of the LC-MS stage of the process. Currently, control limits are not defined for the mass spectrometry component of the affinity isolation experiment. Therefore, trial control limits can be established

Please cite this article in press as: Sharp, J.L., et al., Statistically appraising process quality of affinity isolation experiments. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.011

from the data for these three control charts. When one or more of these control charts indicate an out-of-control signal, then the analysts will need to investigate assignable causes for shifts in the process mean or variability. The instability of the process could be due to many things: a column change in the mass spectrometer, a new batch of the standard protein mixture, or a different mass spectrometer operator.

While the demonstration presented here was performed retrospectively, it would be desirable to implement SQC in a time-frame that would provide feedback for quickly identifying and correcting problems with the LC-MS instrumentation. Currently, peptide identifications are performed via Sequest in a batch mode after LC-MS data acquisition has been completed for a set of standard and experimental samples. These peptide identifications constitute the input data for the SQC analysis presented in this paper, thus precluding a "real time" result. One way to shorten the time required for feedback would be to perform the Sequest analysis on raw LC-MS data from standard samples immediately after acquisition. The delay currently imposed by the Sequest identification step could possibly be eliminated by designing a SQC method based on an attribute of the LC-MS raw data, such as areas under selected chromatographic peaks corresponding to peptides in the standard mixture.

After the cause for the shift in the process mean or variance has been identified and the magnitude of the shift is estimated (Montgomery, 2005), the appropriate adjustment to the process should be made. In addition, researchers will need to specify the in-control process average run length (ARL_0) before a false out-of-control signal occurs and the out-of-control process average run length (ARL_1) before a true out-of-control signal occurs, so that the cusum parameter values h and k can be established.

Acknowledgements

This research was funded by the Genomics: GTL program and the U.S. Department of Energy Office of Advanced Scientific Computing Research under contract 47901, and by the Office of Biological and Environmental Research under contracts 41966 and 43930. Pacific Northwest National Laboratory is a multi-program national laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05-76RL01830. Oak Ridge National Laboratory is managed and operated by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

We wish to thank Trish K. Lankford at Oak Ridge National Laboratory for her expert technical assistance.

References

Berggard, T., Linse, S., James, P., 2007. Methods for the detection and analysis of protein–protein interactions. Proteomics 7, 2833–2842.

Elias, J.E., Gygi, S.P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods 4, 207–214.

Eng, J., Yates, J., 1999. Sequest. http://fields.scripps.edu/sequest/start.html.

Eng, J.K., McCormack, A.L., Yates III, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acide sequences in a protein database. Journal of the American Society for Mass Spectrometry 5 (11), 976–989.

Hawkins, D.M., Olwell, D.H., 1998. Cumulative Sum Charts and Charting for Quality Iimprovement. Springer-Verlag New York, Inc., New York, NY.

Lucas, J.M., 1985. Counted data cusum's. Technometrics 27 (2), 129–144.

Montgomery, D.C., 2005. Introduction to Statistical Quality Control, fifth ed. John Wiley & Sons, Inc.

SAS Institute Inc., 1995. SAS/QC® Software: Usage and Reference, Version 6, Volumes 1 and 2, first ed. SAS Institute Inc., Cary, NC.

SAS Institute Inc., 2000–2004. SAS 9.1.3 Help and Documentation. SAS Institute Inc., Cary, NC.

Smith, J.C., Lamber, J.P., Elisma, F., Figeys, D., 2007. Proteomics in 2005/2006: Developments, applications and challenges. Analytical Chemistry 79, 4325–4343.

Strader, M.B., Tabb, D.L., Hervey, W.J., Pan, C., Hurst, G.B., 2006. Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic–aqueous solven systems. Analytical Chemistry 78 (1), 125–134.

Tabb, D.L., McDonald, W.H., Yates, W.H., 2002. DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. Journal of Proteome Research 1, 21–26.

von Neumann, J., Kent, R.H., Bellinson, H.R., Hart, B.I., 1941. The mean square successive difference. The Annals of Mathematical Statistics 12 (2), 153–162.