Inter-session reproducibility measures for high-throughput data sources

Milos Hauskrecht, PhD, Richard Pelikan, MSc
Computer Science Department, Intelligent Systems Program, Department of Biomedical Informatics, University of Pittsburgh, PA

Abstract

High-throughput biological assays such as micro-arrays and mass spectrometry (MS) have risen as potential clinical tools for disease detection. Multiple potential biomarkers can be rapidly and cheaply evaluated for a large number of patients. Typical research and evaluation studies in these fields have focused primarily on data that were generated from samples in a single data-generation session. However, in the clinical setting, new patients screened by the technology will arrive at different times and data will unavoidably come from multiple data-generation sessions. The understanding and assessment of multi-session effects on data generated by the technology is critical for its application to clinical practice. This paper proposes a methodology for measuring and testing the reproducibility of various aspects of high-throughput data across multiple data-generation sessions. We test and demonstrate the framework on mass-spectrometry data obtained from four different data-generation sessions for the same set of samples.

Introduction

Multiple novel technologies that measure expression levels of genes and complex protein mixtures in the human body offer a great potential for improved detection and understanding of the disease. However, these new technologies are not ideal; the data they generate are subject to various sources of bias and noise[2,3]. Naturally, these affect the quality of signals and their subsequent analyses. The assessment of these influences and their scope is critical for utilization of the technology and data it generates in clinical settings. The goal of this work is to develop an initial set of practical reproducibility measures that can help with this assessment.

Reproducibility refers to our ability to extract the same pattern (signal, statistic) in profile data that are generated by varying (intentionally or unintentionally) some conditions of the data generation process. For example, data may be generated at different times (and different instrument runs), or in different labs. The sample collection and preparation protocols may also vary. This variation can affect the usefulness of the assay readings.

Reproducibility analyses of the data are critical for understanding the utility of the technology and its limits in practice. More specifically, in any practical clinical setting, the data generated by a profiling technology will be (unavoidably) generated in multiple data-generation sessions; the patients and their samples will be analyzed based on the demand and arrival time. Unlike many clinical research studies, the data for these samples cannot be generated all at once during the same data generation session. Hence, the success of the technology and the models developed for it will depend on the reproducibility of the profiles across multiple data generation sessions.

The practical concerns related to the application of these technologies prompts us to acknowledge and study new sources of noise and bias encountered in multi-session settings and to assess their potential detrimental effects. To address these needs, we propose a methodology for testing the reproducibility of various aspects of high-throughput data across multiple data-generation sessions. We define methods for four different reproducibility tests: the presence of the inter-session noise, similarity of profiles across multiple sessions, pattern reproducibility, and generalizability of mixed-session models. We demonstrate these methods by analyzing MS proteomic profiling technology using data obtained from four different data-generation sessions for samples from a lung cancer study.

Methodology

Our evaluation framework is based on the assumption that data for the same set of samples are generated multiple times. We refer to these data generation events as to *sessions*. An ideal technology would yield the same measurement for the same sample across multiple sessions. However, in reality the repeated measurements for the same sample often differ. It is the amount of variation that is critical for the reproducibility assessment. If the data differ widely, it is unlikely the technology and its models will generalize very well to future sessions.

In this work, we propose four methods for measuring the reproducibility of data from multiple data generation sessions:

(1) the presence of intersession noise,

(2) the similarity of inter-session profiles for the same sample,

(3) reproducibility of aggregate patterns,

(4) generalizability of intersession models.

In the following we briefly describe the aims and the basics of these methods.

Test 1: Intersession noise

The fact that profiles for the same set of samples (patients) may differ from session to session makes us wonder whether there is any noise or bias introduced by combining profiles from multiple data-generation session. The answer to this question is not straightforward. The data generated by bioinformatics technologies are inherently noisy, so all of the observed noise may be due to this variation. Our goal is to check if there is any detectable intersession noise component that complicates the problem. We define a paired t-test that focuses on the following null hypothesis.

Hypothesis: Multiple sessions do not add noise to the spectra.

Method:

- Select randomly 1000 pairs of samples (patients).
- Calculate a distance (defined by some distance metric) in between the two spectra for each sample pair (a,b) such that:
  (1) both a and b are from the same session;
  (2) a is from the same session as (1) and be is from a different (randomly chosen) session.
- Show that the mean of distances for (1) is (significantly) better (smaller) than for (2) using a paired t-test, proving that the spectra differences within the same session are smaller and that mixing sessions may add additional noise.

Test 2: Profile similarity

If the intersession noise is present one becomes concerned whether the profiles that originate from the same sample but from different sessions are close enough to warrant some cross-session reproducibility. Clearly, if the intersession noise is large, profiles for individual patients may become very different across sessions and basically undistinguishable from other profiles. Our second method aims to test our ability to distinguish (or not to distinguish) the profiles from the same sample with respect to profiles from other samples across different sessions. Once again we use a paired t-test with the following null hypothesis.

Hypothesis: The spectra for the same patient are undistinguishable across sessions from spectra for other patients.

Method:

- Select randomly 1000 pairs of sessions.
- Calculate the distance in between spectra for two sessions (a,b) such that:
  (a) a and b are for the same patient,
  (b) a is from the same patient as (1) and b is from a different (randomly chosen) patient.
- Show that the mean of distances for (1) is (significantly) better (smaller) than for (2) using a paired t-test, proving that the spectra for the same sample but different sessions are on average more similar to each other than to other samples.

Test 3: Pattern reproducibility

Evaluating profile similarity across sessions helps us assure us of the basic consistency (reproducibility) of profiles with respect to samples they represent. However, data obtained by the technologies are used in variety of interpretive analyses and our concern is that useful statistical patterns may be lost or at least significantly compromised when multiple session data are used. For example. case/control studies typically perform differential and classification analyses. Such analyses are done over complete datasets and the pattern found is an aggregate signal obtained from multiple profiles. To understand the reproducibility of such aggregate patterns we propose a test that measures the quality of a pattern found in data from one session versus the average quality of a pattern mined from data constructed from multiple sessions. Since there are many ways to build combined (multi-session) datasets, instead of enumerating them, we propose to generate mixed session datasets randomly and use their empirical average in the comparison. The loss of the signal due to multi-session data mixing would yield results that are lower on average than those for pure session signals. Conversely the gain would results in better than average results. The overall effect of multi-session mixing can be assessed by averaging the results for individual sessions.

Hypothesis: The pattern is stronger (weaker) in the single session data than in the multi-session data.

Method:

- Generate 1000 random mixed-session profile datasets such that there is one entry per sample.

A profile for the sample is selected randomly from multiple sessions.

- Calculate pattern statistics for: (1) pure sessions and (2) 1000 random mixed-session datasets.
- Assess differences between the pattern statistic for (1) and the average of the statistic over (2).

Test 4: Model generalizability

In the 'ideal' analytical setup for patient profiling studies, a predictive model is trained and evaluated on data from the same session. It experiences only the within-session noise and does not account for potential inter-session noise, should it be re-used for future prediction of profiles. However, in the practical setting of clinical screening, new samples may be processed on-the-fly, each at a different time and therefore experiencing unanticipated amounts of inter-session variability. Concerns about this inter-session reproducibility are related primarily to concerns over generalizability of predictive models that are extracted from past data sessions to profiles obtained in the future. We propose to analyze this aspect of the problem by learning predictive models that are tested on profiles from one target (test) session and trained on the profiles from the remaining (training) sessions and by comparing them to the 'ideal' model trained and tested on the profiles from the same session. The goal of the analysis is to assess quantitatively the performance drop due to this more realistic model learning and testing process and contrast it to the ideal one-session process. Once again, instead of generating all possible multi-session datasets we use randomization to generate empirical averages of observed performance statistics.

Method:

- Generate disjoint training and test sample sets (multiple train/test splits are possible).
- Select one of the sessions as the target (test) session.
- For every training sample set generate 1000 random mixed-session datasets such that there is one entry per sample. A profile for the sample is selected randomly from among multiple available sessions, but excluding the target (test) session.
- Learn classification models for:

  (1) 1000 randomized mixed-session training datasets

  (2) profiles from the same session as the test session

- Compare models learned in (1) and (2) and assess their statistical differences.

This will let us compare the average future performance of mixed-session models to the 'ideal' model for the fixed session only. If necessary, the 'global' performance can be assessed by varying the target test session and/or by averaging the results across all multiple train/test sessions.

Experimental evaluation

To demonstrate the applicability of our reproducibility framework we study it on the MS proteomic data generated for the study of lung cancer. The data consist of 21 lung cancer and 25 control serum samples that were analyzed four times in separate data generation sessions that occurred in June 2003, February 2004, November 2004 and January 2005. The samples collected were part of a larger lung cancer study conducted at the University of Pittsburgh Cancer Institute (UPCI)[1].

Data preparation. Full sample processing protocols for this study have been previously detailed[1,5]. The raw spectra obtained by MS instrumentation were preprocessed using PDAP – a proteomic data analysis package[4]. The preprocessing steps applied included baseline correction, cube-root variance stabilization, total ion current normalization, smoothing with local Gaussian kernels and quadratic Savitzky-Golay filter, averaging of spectra duplicates generated in every session and peak selection.

Analysis. The goal of this analysis is to assess reproducibility of the MS technology over multiple data generation separated by large amounts of time. Our hope was that the differences in multi-session and 'ideal' single-session settings are relatively small and do not affect the results by a large margin. An initial analysis of this multi-session data was performed and published. The contribution of this work is the formalization of the reproducibility framework and its tests, as well as the inclusion of more refined reproducibility metrics (as compared to the initial work) that let us analyze both local and global aspect of the data and their reproducibility.

Results

Test 1 analyzes the presence of additional noise that may be introduced by multiple data generation sessions. The distance metric used in the experiment is equal to absolute differences in between the two signals. (Figure 1) shows the means of these profile differences for pairs of samples from one-session versus two-session data for peaks in the spectra. In

addition to the peak-by-peak analysis of differences in (Figure 1), we also assessed and compared the differences for complete spectra. In this case, the mean of the fixed session differences is 13.0245 and the mean for inter-session differences is 14.5967. This difference was highly statistically significant. Clearly, the differences for one-session data are smaller than for the two-session data, demonstrating additional inter-session noise.

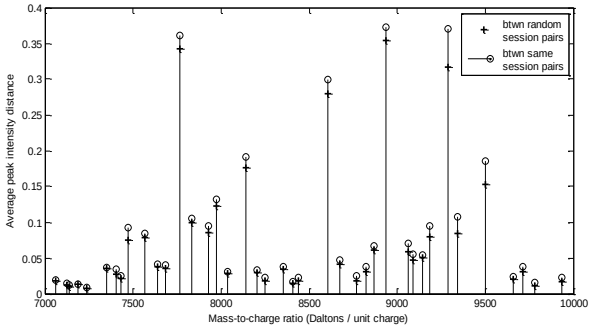

Figure 1. Mean fixed-session and mean inter-session differences for two samples on peaks in the range of 7000 to 10000 Da. The means for fixed session data are labeled by circles; the means for the inter-session data are labeled by crosses.

Test 2. The aim of test 2 was to demonstrate that spectra from the same sample are on average more similar to each other than to profiles for other samples. This test is particularly important if the presence of inter-session noise is confirmed by Test 1. The test gives us hope that the profile differences are preserved. (Figure 2) shows the peak-by-peak analysis of mean differences in between profiles generated for the sample versus mean differences between profiles for two different samples. In summary, out of all peaks the two signals from the same sample were on average closer than signals from two different samples in 95% of peaks. In addition to peak-based analysis we analyzed the differences for all peaks in profiles. The mean differences in between the profiles for the same sample across two sessions were 9.7790 and differences in between two different samples were 14.5967. Once again these mean differences are statistically significantly different, hence demonstrating the similarity of the profiles for the same sample.

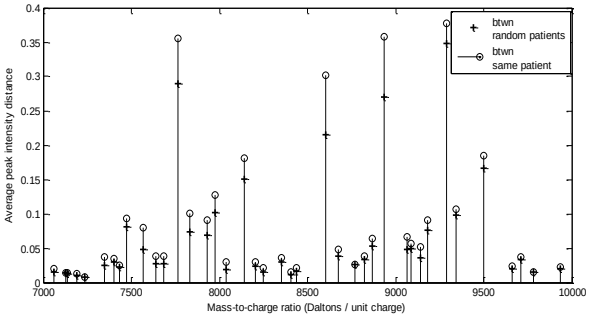

Figure 2. Mean same-sample and mean two-sample inter-session differences on peaks in the range of 7000 to 10000 Da. The means for the same-sample data are labeled by circles; the means for the two-sample data are labeled by crosses.

Test 3 allows the comparison of patterns one can mine in mixed-session and single session datasets and evaluates the influence of the inter-session variability on these patterns. Figure 3 compares differential expression score for peaks in the MS profile for four single-session datasets (labeled by crosses) to the mean score for mixed-session datasets (labeled by circles). The differential expression in the experiment is measured via the Fisher score[6]. (Figure 4) averages the results in (Figure 3) across all peaks. (Table 1) displays the accuracies of multivariate classification models when these are trained on mixed-session versus single session data. The classification models used in the experiment are support vector machines[7]. The results in (Figure 4) and (Table 1) show that there are differences among individual data generation sessions and that some of them carry stronger discriminative signals than others. In particular, our analysis revealed a stronger signal in Session 4 and Session 2 data, while Session 1 results are the weakest. In such a case, multi-session data appear to have averaging effect and helps us to eliminate the influence of very noisy sessions. However, when statistics are averaged over all four sessions the results demonstrate the average loss of a discriminative signal and classification model accuracies in multi-session settings.

Test 4. Table 2 summarizes the results of test 4, which assesses the generalizability of classification models trained on our multi-session data. These models are compared to the 'ideal' setting where a classifier is trained and tested on data from the same session. On average, the advantage naturally falls to the ideal setting models. However the differences are not large which shows a very reasonable generalizability of models built from mixed session data.

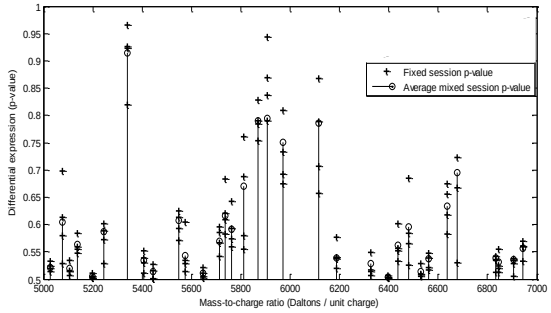

Figure 3. Differential expression for four single-session datasets versus the mean differential expression for mixed-session data. The means for the mixed-session data are labeled by circles; the scores for the single-session data are labeled by crosses.

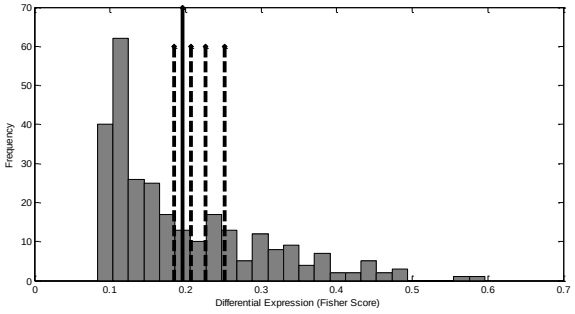

Figure 4. Average differential expression score across all profile peaks. The distribution of scores (mean shown by solid line) for mixed session data is shown and compared to scores obtained for fixed single session datasets (dashed lines).

Chosen sessions	Test 3 Classification
Session #1	84.33%
Session #2	85.50%
Session #3	80.83%
Session #4	87.67%
Mixed sessions average	81.42%

Table 1. Gains and losses in classification accuracy obtained using multivariate patterns.

Conclusion

We have presented methods for assessing the reproducibility of high-throughput biological assay technology over multiple data generation sessions. The results provide an initial assessment of reproducibility that help one to see the benefits and losses from combining data sets generated by the technology. However, we note that reproducibility measures presented in this work make the assessment only for a fixed number of observed data-generation sessions, but they do not provide any guarantees of results under more general multi-session generation conditions. Hence, our work is just a first step in this direction, and more sophisticated statistical reproducibility analyses need to be devised.

Test Session	Mixed Sessions	Ideal Setting
Session #1	79.61%	82.04%
Session #2	87.12%	85.73%
Session #3	75.93%	80.92%
Session #4	85.74%	88.39%

Table 2. Test 4 classification accuracies when training a future-looking model on mixed data sessions versus models trained under the ideal setting.

Acknowledgement

This work was supported in part by Department of Defense grant USAMRAA W81XWH-05-2-0066, NCI grant P50 CA090440-06, and NLM training grant 5 T15 LM007059-20.

References

1. Pelikan R, Bigbee WL, Malehorn D, Lyons-Weiler J, Hauskrecht M. Intersession Reproducibility of Mass Spectrometry Profiles and its Effect on Accuracy of Multivariate Classification Models. Bioinformatics 2007; Aug 30.
2. Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst. 2005;97(4):307-309.
3. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004; 20(5):777-785.
4. Hauskrecht M, Pelikan R, Malehorn DE et al. Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles. Appl Bioinformatics. 2005; 4(4):227-46.
5. Semmes OJ, Feng Z, Adam BL, et al. Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility. Clin Chem. 2005 Jan; 51(1):102-12.
6. Furey TS, Christianini N, Duffy N et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; 16(10):906-914.
7. Vapnik VN. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.