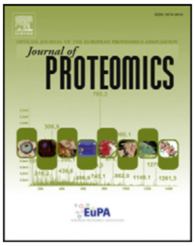
Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



A novel spectral library workflow to enhance protein identifications☆

Haomin Li[a,b,c,1], Nobel C. Zong[a,b,1], Xiangbo Liang[a,b,1], Allen K. Kim[a,b], Jeong Ho Choi[a,b], Ning Deng[c], Ivette Zelaya[a,b], Maggie Lam[a,b], Huilong Duan[c,*], Peipei Ping[a,b,*]

[a]Department of Physiology, UCLA School of Medicine, Los Angeles, CA 90095, USA
[b]Department of Medicine, UCLA School of Medicine, Los Angeles, CA 90095, USA
[c]Department of Biomedical Engineering, Key Laboratory for Biomedical Engineering of Ministry of Education, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Available online 4 February 2013

Keywords:
Error propagation control
Library searching
Noise control
Sliding dot product

ABSTRACT

The innovations in mass spectrometry-based investigations in proteome biology enable systematic characterization of molecular details in pathophysiological phenotypes. However, the process of delineating large-scale raw proteomic datasets into a biological context requires high-throughput data acquisition and processing. A spectral library search engine makes use of previously annotated experimental spectra as references for subsequent spectral analyses. This workflow delivers many advantages, including elevated analytical efficiency and specificity as well as reduced demands in computational capacity. In this study, we created a spectral matching engine to address challenges commonly associated with a library search workflow. Particularly, an improved sliding dot product algorithm, that is robust to systematic drifts of mass measurement in spectra, is introduced. Furthermore, a noise management protocol distinguishes spectra correlation attributed from noise and peptide fragments. It enables elevated separation between target spectral matches and false matches, thereby suppressing the possibility of propagating inaccurate peptide annotations from library spectra to query spectra. Moreover, preservation of original spectra also accommodates user contributions to further enhance the quality of the library. Collectively, this search engine supports reproducible data analyses using curated references, thereby broadening the accessibility of proteomics resources to biomedical investigators.

This article is part of a Special Issue entitled: From protein structures to clinical applications.

1. Introduction

Proteomics investigations in biology and medicine generate an enormous amount of mass spectra daily worldwide [1,2]; specialized bioinformatics platform offers access to biological insights embodied in the raw datasets. In proteomic studies, correlating peptide spectra with their sequences is a vital step towards protein characterization. Such a correlation can be

Abbreviations: DB, dot bias; DP, dot product; NDP, noise dot product; SDP, sliding dot product.

☆ This article is part of a Special Issue entitled: From protein structures to clinical applications.

[*] *Corresponding authors at:* UCLA School of Medicine, 675 Charles E. Young Drive South, MRL Building Suite 1-609, Los Angeles, CA 90095, USA. Tel.: +1 310 267 5624; fax: +1 310 267 5623.

E-mail address: pping@mednet.ucla.edu (P. Ping).

[1] These three authors contributed equally to this work.

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.01.026

established utilizing theoretical spectra, as demonstrated by SEQUEST [3] and Mascot [4], or alternatively, empirical spectra with known peptide identities as a reference [5]. Both approaches exhibit unique strengths and are complementary in nature.

Attributed to its independence from experimental observations, the sequence database search approach has been effective since the past decade, as it catalyzed a wide-acceptance of mass spectrometry-based proteomic investigations [3,6]. However, this popular approach has several inherent limitations. For example, the size of a theoretical spectral dataset derived from a proteome database is quite large and inflates exponentially when post-translational modifications are considered [7]. The emergence of faster scanning mass spectrometers impose further strains on existing bioinformatics pipelines [8]. Therefore, dedicated high-performance computational workstations become a mandate to confer reasonable analytical efficiency and throughput [9,10]. The acquisition and maintenance of such platforms also constitute a significant barrier preventing investigators from embracing proteomics science. Furthermore, the lack of a learning capability of the sequence database search workflow leads to repeated misinterpretations of raw spectra.

The application of empirical spectra as reference for peptide identification addresses challenges associated with a sequence database search workflow [2]. The confined search space of the spectral library focuses on subproteome-of-interest, providing superior specificity and sensitivity [11]. Furthermore, the incorporation of spectra from post-translationally modified peptides leads only to a linear increase in dataset volume. More importantly, the spectral library evolves with iterative contributions from the users. Collectively, the workflow of protein characterization with a spectral library reduces demands on both computational hardware and time, minimizing the barrier of proteomic investigations for biologists and clinicians.

Despite many apparent advantages, enthusiasm towards the application of spectral library has been growing at a rather relaxed rate. The reluctance originates from two related processes: the challenges affiliated with library construction and the accuracy of spectral matching. The quality of spectra is defined by the technical proficiency of current mass spectrometry instrumentations. During the process of library construction, the threshold for spectral quality inversely affects its coverage. Admitting spectra of moderate quality may aid the expansion of the coverage at the expense of compromising analytical accuracy. Falsely annotated spectra may contaminate a spectral library and instigate error propagation [6,12,13]. Developing computational models to circumvent these issues in a library-directed workflow can significantly improve its utility in proteomics investigations.

In this study, we engineered a new workflow for spectral library construction and spectral matching. Unaltered experimental spectra were collected and compiled into a reference library. In parallel, a noise control spectrum was generated for each reference spectrum collected in this library. Accordingly, this data structure enables a spectral matching algorithm to differentiate peptide fragments and noise in the user spectra, enhancing the accuracy of analyses. This integrated bioinformatics platform demonstrated sound performance in protein characterization through parallel benchmark tests using independent datasets from various instrumentations. Taken together, this workflow mobilizes mass spectra accumulated by the proteomics community, which serves as a foundation for future investigations.

2. Materials and methods

2.1. *Collection of raw mass spectra and spectral analyses via sequence database search*

Peptide spectra previously collected from purified murine 20S proteasome complexes were selected to construct a spectral library [14–16]. The purification procedure [16] isolated both proteasome subunits and their interacting partners [17]. A total of 14 biological sample replicates were analyzed, totaling 190 LC–MS/MS runs. These data have been integrated to construct a spectral library.

Theoretical sequence database search was conducted with a SEQUEST search engine (BioWorks cluster, V3.3) on a Beowulf cluster [18]. The IPI mouse database (V3.47, 55,298 entries) was selected as the reference [19]. Endoproteolytic specificity was set to be semi-tryptic with a maximum of two missed cleavages. 50 ppm was assigned for deviations in peptide precursor masses measured with an LTQ-Orbitrap, and 2 amu was assigned for deviations in peptide precursor masses measured by an LTQ instrument. Allowance for fragment mass deviation was set at 1 amu for both cases. Carbamidomethylation on cysteine residues (+57.02146 Da) was set as the static modification; whereas oxidation of methionine residues (+15.9949 Da) and acetylation of protein N-termini (+42.0106 Da) were set as two differential modifications. Scaffold (Proteome Software, V2.0) [20] was applied to filter results at confidence thresholds of 95% for peptides and 99% for proteins. Only proteins with two or more distinct peptides detected were considered as positive identifications.

2.2. *Design of a noise control spectra system for the library*

To integrate noise information into our spectral library, we created a reference spectral component with negative controls. For every peptide, multiple spectra may exist in the spectral library due to variations in their charge states (e.g., +2 or +3) or modification status (e.g., methionine oxidation). Two spectra were implemented for each combination of charge state and modification status. The first spectrum was an unaltered experimental spectrum, selected according to the highest Xcorr score computed by a sequence database search engine. In parallel, a second spectrum containing only noise peaks was constructed from the first spectrum in silico, which is referred to as the 'noise control spectrum' hereafter. Based on the sequence of the assigned peptide, the noise control spectrum was constructed by removing peaks within a ±2 Th window of a-, b- and y- ions and their common neutral loss ($-H_2O$, $-NH_3$) ions from the representative spectrum. Accordingly, an entire collection of representative spectra and their corresponding noise control spectra were created and stored in pairs. A total of 2476 spectral pairs were compiled in the murine 20S proteasome spectral library. This effort