JOURNAL OF PROTEOMICS 99 (2014) 123–137

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism

Daniel J. Stekhoven[a,*,1], Ulrich Omasits[a,b,1], Maxime Quebatte[c], Christoph Dehio[c], Christian H. Ahrens[a,**]

[a]Quantitative Model Organism Proteomics (Q-MOP), Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[b]Institute of Molecular Systems Biology, ETH Zurich, Auguste-Piccard-Hof 1, 8093 Zurich, Switzerland
[c]Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland

ARTICLE INFO

Article history:
Received 12 September 2013
Accepted 15 January 2014
Available online 28 January 2014

Keywords:
Subcellular localization
Experimental proteomics data
Outer membrane proteome
Machine learning
Localization change
Prokaryote

ABSTRACT

Proteomics data provide unique insights into biological systems, including the predominant subcellular localization (SCL) of proteins, which can reveal important clues about their functions. Here we analyzed data of a complete prokaryotic proteome expressed under two conditions mimicking interaction of the emerging pathogen *Bartonella henselae* with its mammalian host. Normalized spectral count data from cytoplasmic, total membrane, inner and outer membrane fractions allowed us to identify the predominant SCL for 82% of the identified proteins. The spectral count proportion of total membrane versus cytoplasmic fractions indicated the propensity of cytoplasmic proteins to co-fractionate with the inner membrane, and enabled us to distinguish cytoplasmic, peripheral inner membrane and bona fide inner membrane proteins. Principal component analysis and k-nearest neighbor classification training on selected marker proteins or predominantly localized proteins, allowed us to determine an extensive catalog of at least 74 expressed outer membrane proteins, and to extend the SCL assignment to 94% of the identified proteins, including 18% where *in silico* methods gave no prediction. Suitable experimental proteomics data combined with straightforward computational approaches can thus identify the predominant SCL on a proteome-wide scale. Finally, we present a conceptual approach to identify proteins potentially changing their SCL in a condition-dependent fashion.

Biological significance
The work presented here describes the first prokaryotic proteome-wide subcellular localization (SCL) dataset for the emerging pathogen *B. henselae* (*Bhen*). The study indicates that suitable subcellular fractionation experiments combined with straight-forward computational analysis approaches assessing the proportion of spectral counts observed in different subcellular fractions are powerful for determining the predominant SCL of a large percentage of the experimentally observed proteins. This includes numerous cases where *in silico*

* *Correspondence to:* D.J. Stekhoven, Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland.
** *Correspondence to:* C.H. Ahrens, Agroscope, Institute for Plant Production Sciences, Research Group Molecular Diagnostics, Genomics and Bioinformatics, Schloss 1, CH-8820 Wädenswil, Switzerland. Tel.: +41 44 783 6114.
E-mail addresses: stekhoven@quantik.ch (D.J. Stekhoven), christian.ahrens@agroscope.admin.ch (C.H. Ahrens).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.jprot.2014.01.015

prediction methods do not provide any prediction. Avoiding a treatment with harsh conditions, cytoplasmic proteins tend to co-fractionate with proteins of the inner membrane fraction, indicative of close functional interactions. The spectral count proportion (SCP) of total membrane versus cytoplasmic fractions allowed us to obtain a good indication about the relative proximity of individual protein complex members to the inner membrane. Using principal component analysis and k-nearest neighbor approaches, we were able to extend the percentage of proteins with a predominant experimental localization to over 90% of all expressed proteins and identified a set of at least 74 outer membrane (OM) proteins. In general, OM proteins represent a rich source of candidates for the development of urgently needed new therapeutics in combat of resurgence of infectious disease and multi-drug resistant bacteria. Finally, by comparing the data from two infection biology relevant conditions, we conceptually explore methods to identify and visualize potential candidates that may partially change their SCL in these different conditions. The data are made available to researchers as a SCL compendium for *Bhen* and as an assistance in further improving *in silico* SCL prediction algorithms.

1. Introduction

In addition to information about the relative abundance of proteins, their modifications and interaction partners, proteomics data contribute yet another important and distinctive type of information: the subcellular localization (SCL) of proteins. Such knowledge can provide initial or additional clues about a protein's function [1,2], and often represents the first experimental evidence that a predicted ORF lacking any homology or functional annotation (frequently called "hypothetical ORF") is a true protein-coding gene [3]. SCL data can furthermore benefit from adding constraints *via* integration of protein–protein interaction [4] and metabolic reaction networks [5], which are becoming increasingly important for systems biology. Finally, the experimental identification of a protein in an unexpected SCL is often a first indicator that a protein may carry out an additional function beyond a well-established one, a phenomenon referred to as moonlighting [6], which has also been linked with bacterial virulence [7].

Fueled by large advances in proteomics [8,9], an ever-increasing coverage of proteomes expressed in specific states or conditions has been achieved [10]. With the notable exception of landmark studies in yeast, which had determined the predominant SCL of 70% of the predicted baker's yeast proteins using GFP tags [1], and 90% of the encoded fission yeast ORFs using YFP tags [2], experimental SCL data have been less comprehensive. Nevertheless, in eukaryotes, the enrichment of specialized organelles or subcellular sub-structures has greatly helped in deciphering their protein make-up and gaining relevant functional insights, *e.g.* for mitochondria [11], and the nucleolus [12]. In addition, excellent software packages have been developed that infer the predominant localization of eukaryotic proteins based on experimental expression data [13,14].

Due to their smaller complexity, prokaryotic proteomes hold particular value for systems-biology motivated studies including the analysis of complete membrane proteomes. In light of the resurgence of infectious diseases, such studies are urgently needed to provide novel targets for diagnostics and therapy of pathogens [15]. Accordingly, SCL analyses in prokaryotes had focused on the important outer membrane proteome [16–19] which holds the potential to identify important virulence factors, proteins raising an immune response in the affected host [20,21], as well as novel targets for therapeutic intervention. Other studies have described the inner membrane proteome, [22,23] peripheral proteins associated with the inner membrane [24,25], the periplasmic proteome [26,27], or on analyzing experimental evidence for several different SCLs [28–30], or have explored combinations of experimental data and *in silico* predictions [31].

Owing to the more limited coverage of experimental methods, *in silico* tools for prediction of the SCL of proteins became popular, which often offer a first valuable starting point. They include LOCkey, one of the early tools developed by the group of Burkhard Rost [32], which relies on sequence homology search, extraction of annotation and text mining to identify predominant SCL, as well as a host of later prediction software like Proteome Analyst v2.0 [33], Cello v2.5 [34], PSORTb v2.0 [35], SherLoc [36], PSORTb v3.0 [37], and LocTree2 [38].

Recently, we reported a strategy to describe complete expressed prokaryotic proteomes [39], and applied it to the emerging pathogen *Bartonella henselae* (*Bhen*), a Gram-negative model for host–pathogen interaction. When grown in two states that mimic conditions encountered in different hosts (Fig. 1a), a total of 1250 proteins were identified, *i.e.* more than 85% of all 1467 distinct *in silico* annotated *Bhen* proteins.

This dataset represented an ideal opportunity to i) test whether experimental proteomics data from different subcellular fractions combined with straightforward computational methods and biological domain expert knowledge can be used to determine the predominant SCL for a large percentage of the identified proteins, to ii) explore how well the *in silico* predictions correlate to the experimental SCL, and to iii) assess for how many proteins we could provide an SCL for the first time. Furthermore, we wanted to explore whether the spectral count proportion (SCP) observed in different fractions could provide hints about the proximity of cytoplasmic proteins to the inner membrane. It has become increasingly clear that under physiological conditions many cytoplasmic proteins are closely attached to inner membrane proteins [25], a fact that is not captured by *in silico* prediction methods, but with important implications regarding protein function. Finally, we wanted to generate a basic computational approach to capture a comprehensive set of OM proteins and, although not being able to draw on replicates (which is not possible for the complete proteome dataset studied here) to explore visual approaches to identify proteins potentially changing their SCL in different conditions.