analytical chemistry

Editors' Highlight

pubs.acs.org/ac

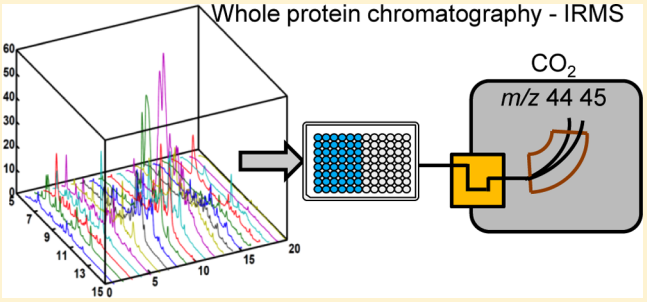
Protein Stable Isotope Fingerprinting: Multidimensional Protein Chromatography Coupled to Stable Isotope-Ratio Mass Spectrometry

Wiebke Mohr,[†] Tiantian Tang, Sarah R. Sattin, Roderick J. Bovee,[‖] and Ann Pearson*

Department of Earth and Planetary Sciences, Harvard University, 20 Oxford St., Cambridge, Massachusetts 02138, United States

Supporting Information

ABSTRACT: Protein stable isotope fingerprinting (P-SIF) is a method to measure the carbon isotope ratios of whole proteins separated from complex mixtures, including cultures and environmental samples. The goal of P-SIF is to expose the links between taxonomic identity and metabolic function in microbial ecosystems. To accomplish this, two dimensions of chromatography are used in sequence to resolve a sample containing ca. 5−10 mg of mixed proteins into 960 fractions. Each fraction then is split in two aliquots: The first is digested with trypsin for peptide sequencing, while the second has its ratio of $^{13}C/^{12}C$ (value of $\delta^{13}C$) measured in triplicate using an isotope-ratio mass spectrometer interfaced with a spooling wire microcombustion device. Data from cultured species show that bacteria have a narrow distribution of protein $\delta^{13}C$ values within individual taxa (±0.7−1.2‰, 1σ). This is moderately larger than the mean precision of the triplicate isotope measurements (±0.5‰, 1σ) and may reflect heterogeneous distribution of ^{13}C among the amino acids. When cells from different species are mixed together prior to protein extraction and separation, the results can predict accurately (to within ±1σ) the $\delta^{13}C$ values of the original taxa. The number of data points required for this endmember prediction is ≥20/taxon, yielding a theoretical resolution of ca. 10 taxonomic units/sample. Such resolution should be useful to determine the overall trophic breadth of mixed microbial ecosystems. Although we utilize P-SIF to measure natural isotope ratios, it also could be combined with experiments that incorporate stable isotope labeling.



It has long been a challenge to study the *in situ* functions of diverse microbial communities.[1−3] Several approaches use measurements of stable isotopes, both at natural levels and by selective isotopic enrichment, to link taxa to their presumed metabolisms. Such techniques include high-resolution analysis of single cells, e.g., secondary ion mass spectrometry (SIMS);[4−6] as well as molecular methods that measure the incorporation of specific substrates into DNA, RNA, or proteins via stable isotope probing (SIP).[7−10] Here, we present protein stable isotope fingerprinting (P-SIF), a new method for measuring natural-abundance stable isotope ratios of proteins extracted from multispecies mixtures. P-SIF has the potential to link metabolic processes to taxonomic identity without the introduction of exogenous labels, or alternatively, to minimize the concentrations of such labels and the attendant incubation times.

Proteins generally account for the majority of total microbial cell mass by dry weight, and their $^{13}C/^{12}C$ carbon isotope ratios reflect the carbon source(s) assimilated by the organism.[11] Further enzymatic fractionation also redistributes these isotopes intracellularly. Such biochemical fractionations can be used as natural metabolic signatures or "fingerprints".[11−13] As a proof of concept, we show that we can distinguish the $^{13}C/^{12}C$ ratios (values of $\delta^{13}C$, here shortened to δ) of proteins extracted from a mixture of two photosynthetic species, one grown on atmospheric CO_2 and one grown on fossil fuel-derived CO_2, and assign the proteins back to their respective sources based on both their isotope ratios as well as their sequences.

Our strategy employs top-down, rather than bottom-up proteomics; i.e., it begins with prefractionation of undigested proteins. Modeled after some recent examples,[14,15] we separate whole proteins using sequential strong anion exchange (SAX) and reverse phase (RP) high performance liquid chromatography (HPLC). The resulting fractions are split, and one aliquot is dried and rehydrated for spooling wire microcombustion (SWiM)[16] isotope ratio mass spectrometry (IRMS),[17,18] while the other is digested to yield a peptide mixture of relatively low complexity for rapid sequencing and taxonomic identification. An advantage of this top-down approach is its compatibility with the sample requirements of SWiM-IRMS. In applications of continuous-flow IRMS, CO_2 is generated by quantitative combustion of the analyte, usually coupled to gas chromatography,[19] but more recently in tandem with liquid chromatography using SWiM or related interfaces.[20−22] SWiM also is useful as a nanocombustion device for

Received: June 29, 2014
Accepted: August 14, 2014
Published: August 14, 2014

   dx.doi.org/10.1021/ac502494b | *Anal. Chem.* 2014, 86, 8514−8520

samples that are prepared as discrete aliquots, either by nanosamplers,[23] by probe-target capture,[24,25] or by preparative chromatography with fraction collection in volatile solvents (this work).

EXPERIMENTAL SECTION AND RESULTS

Cultures and Extracts. *Allochromatium vinosum* DSM180 and *Synechocystis* sp. PCC6803 were grown as described in the Supporting Information and in previous work.[26] Their total biomass yielded δ values of -59.9 ± 0.1‰ and -29.3 ± 0.1‰, corresponding to values of $\alpha_{CO2\text{-biomass}}$ of 1.024 and 1.022 relative to their growth on tank CO_2 and air, respectively (Table 1).

Table 1. Statistics for P-SIF Results Obtained from a Mixture of *Synechocystis* sp. PCC6803 and *A. vinosum* DSM180

P-SIF protocol: total fractions	960
QTOF-MS/MS: Syn + Allo Mix	
fractions with detectable protein	300
total proteins detected	624
total unique proteins	213
mean proteins/fraction	2.2
SWiM-IRMS: Syn + Allo Mix	
predicted # $\delta^{13}C$ values (δ'_m)	265
actual # $\delta^{13}C$ values (δ'_m)	210
predicted data yield	28%
actual data yield	22%
mean σ_δ from standards	±0.50‰
median σ_δ from standards	±0.35‰
mean σ_δ from samples	±0.54‰
median σ_δ from samples	±0.40‰
Synechocystis sp. PCC6803	
biomass (EA-IRMS)[a]	-29.3 ± 0.1‰
proteins (P-SIF), $n = 88$	-27.6 ± 0.7‰
Allochromatium vinosum DSM180	
biomass (EA-IRMS)[a]	-59.9 ± 0.1‰
proteins (P-SIF), $n = 79$	-58.0 ± 1.2‰
Endmember δ Predictions from P-SIF	
Synechocystis sp. PCC6803	−27.7‰ (CI 0.05)[b]
A. vinosum DSM180	−56.9‰ (CI 0.05)[b]

[a]Measured by elemental analyzer (EA)-IRMS. [b]Deming regression confidence interval (CI).

To begin the P-SIF process, approximately equal quantities (0.1−0.2 g) of frozen, pelleted biomass of each culture were pooled before extracting with 1 mL of B-PER Protein Extraction Reagent (Thermo Scientific) and 0.5 mL of 0.1 mm zirconia beads. Samples are broken using a microbead beater (6 × 20 s, with 1 min at 0 °C between each interval) and then incubated at 0 °C for 5 min. After centrifuging for 15 min at 16 000*g* (4 °C), the supernatant is precipitated in acetone (≥5:1 acetone/extract) at 0 °C for 1.5 h and centrifuged (8000*g* × 15 min) to pellet the proteins. Air-dried pellets are dissolved in 100 mM NH_4HCO_3, containing 5 mM dithiothreitol (DTT), 5% isopropanol, and 200 mM glycine and are used immediately for chromatography.

Multidimensional Protein Chromatography. All chromatography is performed on an Agilent 1100 series HPLC with absorbance detection at 280 and 409 nm (see additional methods in the Supporting Information). For the first dimension, 5−10 mg of mixed proteins are separated on a PL-SAX column (4.6 × 50 mm, 8 μm particle size) with a constant solvent flow of 1.0 mL min^{-1}, a temperature of 50 °C, and a gradient from 100% solvent A to 100% solvent B (Table S1, Supporting Information), where A and B are 50 mM and 1 M NH_4HCO_3 (both pH 9.0), respectively. Twenty 1.0 mL fractions are collected between 5 and 25 min in a fraction collector maintained at 10 °C. Fractions are either processed in the second chromatographic dimension within 4 h or stored at −80 °C for later analysis.

The second, RP-HPLC dimension uses a Poroshell 300SB-C_3 column (2.1 × 75 mm, 5 μm particles) with a solvent flow of 0.65 mL min^{-1}, temperature of 65 °C, and a gradient from 97% C to 100% D (Table S1, Supporting Information), where C is H_2O and D is 1:1 isopropanol/acetonitrile, both containing 3% formic acid. Each of the SAX fractions is separated into a 96-well plate; 48 fractions of 0.135 mL are collected between 5 and 15 min, and the remaining wells are reserved for the addition of exogenous protein and amino acid standards and process blanks (Table S2, Supporting Information). Each plate is subsampled immediately into new plates for fluorescent protein quantification (5 μL for NanoOrange; Invitrogen) and for tryptic digestion and sequencing (40 μL). The remaining 90 μL is reserved in the original plates for SWiM-IRMS. All plates are stored at −80 °C until analysis.

The chromatographic peak capacity of the SAX and RP dimensions can be calculated by fitting Gaussian functions to the chromatograms (Figure 1; Figure S1, Supporting Information). Results for *Synechocystis*, *A. vinosum*, and a mixture of the two species show that on average the SAX dimension distinguishes 30 peaks (range of 22−35; Table S3, Supporting Information). This resolution is 50% higher than the number of SAX fractions currently collected, suggesting that in future investigations the sampling frequency should be increased. Analysis of each SAX fraction by SDS-PAGE gel electrophoresis shows significant complexity, but each also has distinct features (Figure S2, Supporting Information).

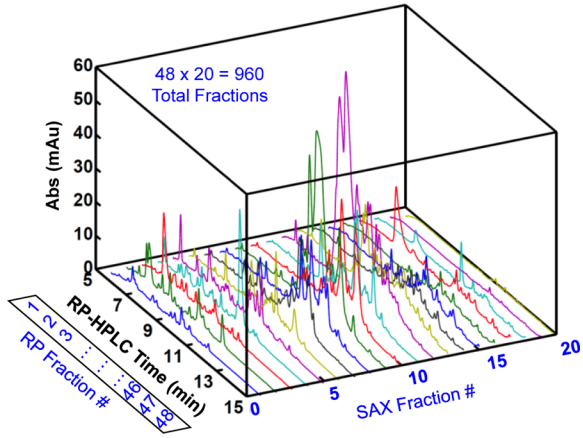


Figure 1. Protein extract obtained from a mixture of *A. vinosum* and *Synechocystis* cells, separated by SAX chromatography followed by RP chromatography. Each of the 20 fractions of the SAX dimension shows both the complexity within, and also the differences between, the adjacent fractions. Each of the RP chromatograms is divided into 48 equal time slices for collection in 96-well plates. For additional comparison of the individual RP chromatograms, see Figure S5, Supporting Information.

RP chromatograms of the mixed *Synechocystis* + *A. vinosum* sample show a range of amplitudes and continue to have significant overlap between the individual proteins (Figure 1;

dx.doi.org/10.1021/ac502494b | *Anal. Chem.* 2014, 86, 8514−8520

Figure S1, Supporting Information). Most runs appear to contain a diversity of proteins well in excess of the resolution, and the resolution depends strongly on the concentrations of individual proteins. Using known protein standards, as well as the samples from the bacterial cultures, we calculate that the peak capacity of this dimension ranges from <30 to >40 peaks run^{-1} (range of 32−54; average 41; Table S3, Supporting Information). Most of the individual peaks from bacterial fractions contain <10 μg of total protein, based on their peak widths and amplitudes when compared to standards. Chromatograms of all *Synechocystis*, *A. vinosum*, and mixed sample RP fractions are shown in Figures S3, S4, and S5, Supporting Information, respectively.

The theoretical average resolving power for our 2D scheme is therefore 30 × 41 = 1230 proteins, which we collect as 960 discrete fractions (20 × 48, Figure 1), or 1440 fractions in future studies (30 × 48). Such resolution is similar to the 1000−2000 proteins maximally resolvable by 2D gel electrophoresis,[27,28] while having the dual advantages of accommodating a larger initial quantity of protein (ca. 5−10 mg) and yielding samples directly in a volatile solvent mixture.

Peptide Sequencing by QTOF-MS/MS. Plates for tryptic digestion are prepared and sequenced as detailed in the Supporting Information. Peptides are identified by capillary LC-MS/MS using an Agilent 1200 Series HPLC equipped with a Kinetex C18 column (2.1 mm × 100 mm, 2.6 μm particles) and an Agilent 6520 quadrupole time-of-flight mass spectrometer (QTOF-MS/MS).

We sequenced all 768 wells from SAX fractions 1−16 of the mixed *Synechocystis* + *A. vinosum* sample. Of these, 300 wells had detectable peptide signal (>ca. 10^3 mean peptide counts) that matched one or more proteins from either species; 624 protein hits were identified (Table 1; Figure S6a, Supporting Information). Some overlap of proteins is observed between adjacent wells, such that the number of unique proteins is 213 (59 from *Synechocystis* and 154 from *A. vinosum*). Protein identifications from SAX fractions SAX.03 and SAX.09 are shown as examples (Figure 2; Figure S7, Supporting Information). The mean number of proteins detected in individual wells is 2.2 (range, 1−10) (Table 1). This number does not scale strictly with peptide signal intensity; i.e., there are many instances of 1−2 proteins identified at >5 × 10^5 mean counts, presumably indicating abundant but relatively pure proteins in these fractions (Figure S8, Supporting Information). Conversely, there is poor detection of multiple proteins at lower signal intensity (no instances of >2 proteins at <5 × 10^4 mean counts), suggesting that low-abundance proteins are being systematically under-detected due to limitations in sensitivity of the QTOF-MS/MS sequencing. The QTOF-MS/MS signal moderately correlates with integrated HPLC absorbance at 280 nm (A_{280}) (R^2 = 0.4) and with CO_2 combustion yield by SWiM-IRMS (R^2 = 0.4) (Figure S9a,b, Supporting Information). For a list of all detected proteins, see the Supporting Information, Table S4.

Determining Values of δ by Automated SWiM-IRMS. SWiM-IRMS has been described previously.[16,20−23] Briefly, ≤1 μL of analyte solution is deposited on a preoxidized, 0.25 mm diameter Ni wire. The wire moves horizontally into a combustion furnace, where the analyte is combusted quantitatively. The resulting CO_2 is separated from H_2O over a Nafion membrane and admitted via an open split to the IRMS. Our system most closely resembles that of Sessions et al.,[16] but with a few significant modifications: the preoxidation and combustion furnace temperatures have been raised to 950 and 800 °C, respectively; the combustion furnace is configured to admit He at both ends, as in Thomas et al.;[22] and the reactor tube is ceramic, not quartz, with dimensions of 3.2 mm o.d. (1/8 in.), 0.8 mm i.d., and 16.5 cm. In addition, we outfitted the instrument with a LEAP-PAL autosampler with chilled sample drawers (10 °C).

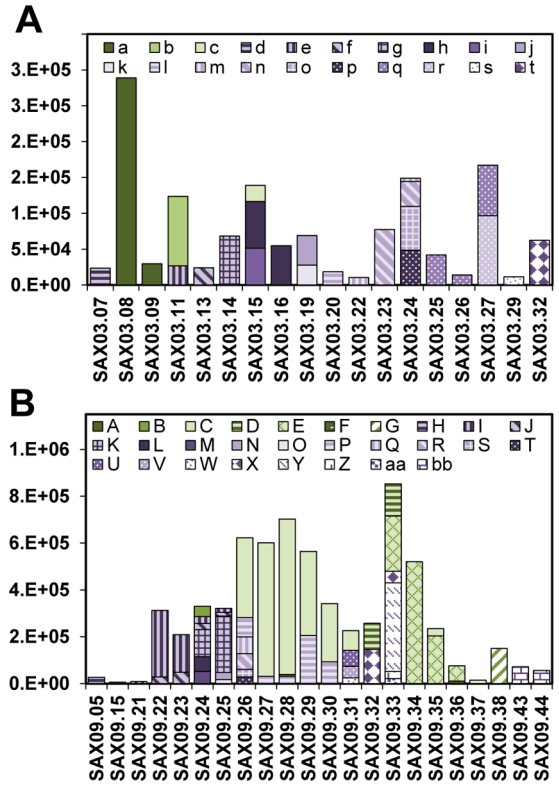


Figure 2. Protein identifications for fractions SAX.03 (A) and SAX.09 (B). Green shaded patterns are hits from *Synechocystis* sp. PCC6803, while purple shaded patterns are hits from *A. vinosum* DSM180. Protein names corresponding to the legend code letters are shown in Figure S7, Supporting Information.

To prepare 96-well plates for IRMS, the RP-HPLC solvents are removed by centrifugal vacuum evaporation (6 h, 37 °C). To aid resolubilization, a mild Fenton oxidation[29] is performed by adding 6 μL each of 0.2 mM $FeCl_3$ and 0.05% H_2O_2, followed by exposure to UV radiation in a biosafety hood for 5−6 min and mild sonication (2 min). Plates sit at 4 °C overnight to complete the rehydration, and IRMS analysis is performed the next day. The Fenton oxidation was optimized using protein standards, and trials were performed to choose the type of 96-well plate having the lowest carbon background (blank), the optimal UV exposure time, and the H_2O_2 concentration (Supporting Information).

To measure values of δ, the autosampler is programmed to inject 0.8 μL of sample onto the SWiM-IRMS wire at an interval of 28 s (Figure S10, Supporting Information) or ca. 100 s for triplicate analysis of a single well, including needle washing. Typically, 30−48 wells are measured per run, using a ca. 3:2 ratio of samples/standards. The most common technical problem is failure of the autosampler needle to leave a drop on the wire (a process that is controlled by surface tension). Because a missed drop may then be mixed with subsequent drops, this compromises the entire triplicate if it occurs during the first drop, or it reduces the number of usable peaks to 1 or 2

dx.doi.org/10.1021/ac502494b | *Anal. Chem.* 2014, 86, 8514−8520

if it occurs during the second or third instances. Two such examples appear in Figure S10, Supporting Information.

The precision and accuracy of reported values of δ is a function of at least three different factors, assessed here by examining our extensive data set for protein and amino acid standards (n = 473 triplicate analyses, or ca. 1400 data points). Following Sessions et al.,[16] our mean precision for CO_2 yield is nearly independent of sample size (mean σ_m/m = 0.09, where m stands for the mass of sample); we use this observation to apply a conservative cutoff (σ_m/m > 0.2) to eliminate ≤5% of the data as outliers (Figure 3A). Standards that pass this initial curation

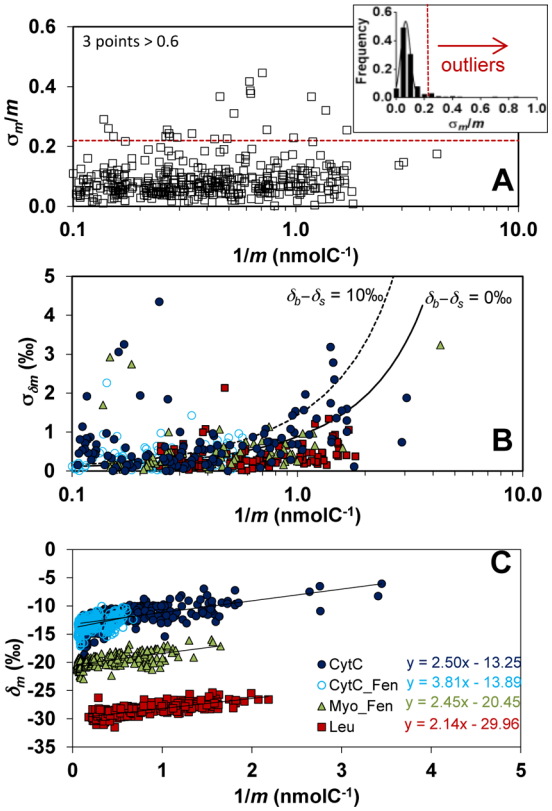


Figure 3. (A) Relative error in CO_2 combustion yields (σ_m/m); each point represents a triplicate analysis of either an amino acid or protein standard. (B) Error in values of δ_m ($\sigma_{\delta m}$), after removing outliers from (A) and sorted by type of standard. (C) Values of δ_m for the same data, after removing those with $\sigma_{\delta m}$ > 2‰; offsets are a function of sample size, m, and are parallel in slope regardless of δ_s.

are examined for a carbon blank of the type that typically is corrected by isotope mass balance: $\delta_m = (b\delta_b + s\delta_s)/m$ where b and s stand for the mass of blank and sample and $m = b + s$. However, when the measurement precision $\sigma_{\delta m}$ is modeled according to eqs (3) and (4) of Sessions et al.[16] (Figure 3B), the data are consistent with our direct measurements of the mass of blank ($b \leq 0.15$ nmolC) and an estimated $\delta_b - \delta_s$ = 0‰, i.e., effectively no blank. The mean precision ($\sigma_{\delta m}$) is ±0.50‰ across all data for standards, while the median precision is ±0.35‰. On the basis of these observations, we conservatively reject all triplicates (both standards and samples) having $\sigma_{\delta m}$ > ± 2.0‰ (ca. 2% of the data) but do not make any further blank corrections. We suspect that the stochastic distribution of values with $\sigma_{\delta m}$ > ± 2.0‰ (and σ_m/m > 0.2) is due to random instances of sample carryover in the SWiM-IRMS system and/or dust falling on the wire.

Additional evidence for lack of a systematic blank comes from examining the absolute values of δ_m obtained for the standards (Figure 3C). Linear functions for δ_m vs. $1/m$ should intersect at $(x, y) = (1/b, \delta_b)$ when measured on standards having different values of δ_s, yet here we find that all lines are nearly parallel, i.e., smaller m correlates with positive bias in δ_m, regardless of the true δ_s. This is strong evidence for an instrument artifact, most likely formation of HCO_2^+ due to residual H_2O in the ionization source.[18] To compensate, we convert measured values of δ_m to corrected values, δ'_m, by applying a linear equation that incorporates the average slope. To further minimize the risk of systematic errors, all data for samples and standards having m < 0.56 nmolC (~350 mV peak amplitude, m/z 44) also are eliminated. Data treated by these approaches are both precise and accurate (±0.5‰) across standards spanning 0.56−18 nmolC injected on the wire (equivalent to 0.02−0.5 μg of protein).

Selection of Protein-Containing Wells for SWiM-IRMS Measurements. It is time-prohibitive and inefficient to measure values of δ for all 960 wells from a single P-SIF separation. To select wells for SWiM-IRMS analysis, their protein content is estimated by two approaches: (i) fluorescent quantification using NanoOrange (Invitrogen) and (ii) integrated spectral absorbance of the RP-HPLC signal at 280 nm (A_{280}). NanoOrange, however, has proved unreliable and will be abandoned in future studies (Figures S6b and S9c, Supporting Information). The results for A_{280}, in contrast, correlate well with the subsequent CO_2 yield on the IRMS (R^2 = 0.6; Figures S6c,d and S9d, Supporting Information). This again suggests there is minimal nonprotein carbon in the individual P-SIF fractions, and yet the scatter of data also indicates that all methods used here for estimating protein content are only semiquantitative.

The concentration distribution ranges from 0 to >10 μg protein/well. This likely reflects both the broad range of intracellular concentrations for individual proteins, as well as the effects of coelution (Figures 1 and 2). In addition, all late-eluting SAX fractions have poor CO_2 yields by SWiM-IRMS relative to predictions from A_{280} (Figures S6c,d and S11, Supporting Information); the reason for this poor performance is still unexplained. The mixture of *Synechocystis* + *A. vinosum* yields data for δ'_m from 22% of all wells, relative to a predicted 28% of wells as calculated from the product of the probabilities in Figure S11, Supporting Information, and the A_{280} data. In total, we obtained 210 values for δ'_m for the two-species mixture (Table 1).

Results for Pure Cultures and the 2-Species Mixture. Figure 4 shows the compiled results for two SAX fractions, SAX.03 and SAX.09, from the *Synechocystis* + *A. vinosum* protein mixture. These represent a low-concentration and a high-concentration fraction, respectively, with dynamic ranges of HPLC absorbances and SWiM-IRMS CO_2 yields >20. Measured values of δ'_m in these examples range from −28.8‰ to −55.9‰. Despite the chromatographic overlap between proteins, which is especially evident in SAX.09 (Figure 4B), the data are consistent with the predicted RP-HPLC resolution of 15 s peak widths. The sharp peak at well 25 of SAX.09 (−55.9‰) is distinguished isotopically from well 27 (−29.7‰). The protein content of well 25 is identified by QTOF-MS/MS as 75% superoxide dismutase, 10% SurA domain-containing protein, and 15% other proteins, all from *A. vinosum*, while well 27 contains 95% phycocyanin

dx.doi.org/10.1021/ac502494b | *Anal. Chem.* 2014, 86, 8514−8520

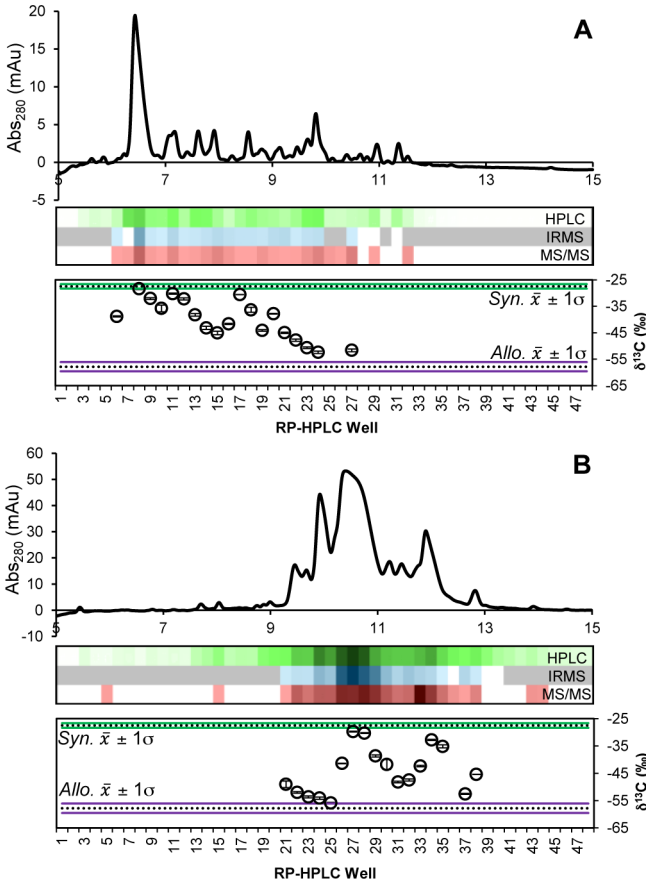


Figure 4. Comparison of protein yield and isotope data obtained from SAX.03 (A) and SAX.09 (B). The upper panel of each shows the RP-HPLC chromatogram (A_{280}); the middle panel shows quantification by three methods (absolute values for color scales in Figure S6, Supporting Information); and the bottom panel shows values of δ'_m for the individual wells. Note the rapid transition from −56‰ to −29‰ across wells 25−27 in (B), representing <30 s of chromatography.

(*Synechocystis*) and 5% ribose 5-phosphate isomerase (*A. vinosum*) (Table S4, Supporting Information).

To interpret the full set of values of δ'_m obtained from the *Synechocystis* + *A. vinosum* protein mixture (Table S5, Supporting Information), the full P-SIF protocol also was performed separately on individual cultures of the two taxa to determine the mean value of δ'_m for each endmember (−27.6‰ for *Synechocystis* proteins and −58.0‰ for *A. vinosum* proteins; Table 1). For reasons of efficiency, we analyzed only enough fractions to yield n = 100 values for each species. After eliminating low-abundance wells (<0.56 nmolC), the actual number of values acquired was 88 and 79, respectively. Gaussian fits to the binned data show a narrow range of isotopic distribution for proteins within the individual species (Figure 5). The wider spread of data for *A. vinosum* (1σ = ± 1.2‰) vs *Synechocystis* (1σ = ± 0.7‰) may be an analytical artifact of working significantly below the isotopic range of our authentic standards (−58‰ for *A. vinosum* vs −29‰ for the most negative standard, leucine), or it may genuinely reflect broader intracellular isotopic heterogeneity in this species. On the basis of these results, we presently regard the 1σ "isotopic breadth" of single bacterial species to be ca. ± 1‰. In future work, more species will be examined to probe the applicability of this number across metabolic types and greater phylogenetic diversity. Regardless, the breadth of values within a taxon is

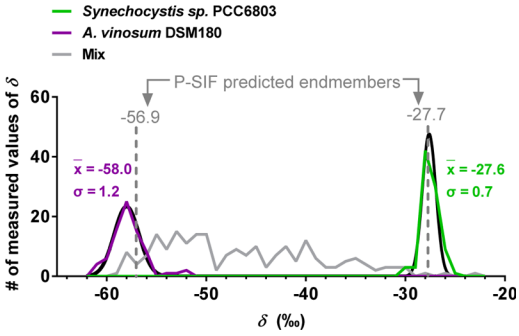


Figure 5. Histogram of values of δ'_m for pure cultures of *Synechocystis* sp. PCC6803 (green line), *A. vinosum* DSM180 (purple line), and a mixture of the two species (solid gray line). The data for individual species are Gaussian (black lines). Regression of all values of δ'_m for the mixture against the fraction of each species as predicted from peptide sequencing yields endmember values for δ_{Allo} and δ_{Syn} (gray dashed lines).

larger than our analytical error (±0.5‰), suggesting that at least some of the breadth is a real signal. Different proportions of ^{13}C-enriched or ^{13}C-depleted amino acids would be expected as a function of protein sequence and may result in different values of δ among proteins, even if synthesized at the same time and from the same pool of cellular metabolites. Values of δ for the individual amino acids within a species span at least 20‰.[30]

For the 2-species mixture, the measured values of δ'_m along with the QTOF-MS/MS peptide signal intensity and protein assignments (Tables S4 and S5, Supporting Information) together constitute an overdetermined set of linear equations of the form:

$$\delta'_{m,i} = (\sum \mathrm{Pep}_{\mathrm{Allo},i} \times \delta_{\mathrm{Allo},i} + \sum \mathrm{Pep}_{\mathrm{Syn},i} \times \delta_{\mathrm{Syn},i})/\sum \mathrm{Pep}_{\mathrm{Total},i} \quad (1)$$

where δ_{Allo} and δ_{Syn} are the mean values for the proteins in the endmember species, ΣPep is the summed QTOF-MS/MS ion counts for peptides, and i is each individual plate well for which both δ'_m and peptides were measured. Among the 210 values of δ'_m, 154 also have data for ΣPep. Because we already know the true values for δ_{Allo} and δ_{Syn} independently (Table 1), the goal here is to evaluate the ability to use the values of $\delta'_{m,i}$ and ΣPep from the 2-species mixture to converge on the correct answers for δ_{Allo} and δ_{Syn} for the individual species. Importantly, this approach to data treatment can be extended to mixed, natural microbial ecosystems, because eq 1 can be a linear sum of any number of variables, each representing a taxonomic division.

Here, the simple two-endmember system can be solved by regression of δ'_m against the fractional abundance of one taxon as predicted from the peptide data. The intercepts at 0% and 100% abundance define the value of δ for each endmember. Indeed, if the δ'_m and ΣPep measurements were perfectly accurate, all that would be needed would be to find all wells containing 0% and 100% *Synechocystis* peptides and confirm that the mean δ values of those bins matched the expected values for δ_{Allo} and δ_{Syn}, respectively. An examination of the raw data shows immediately why this will not work. When the ΣPep data are viewed in rank-order of signal intensity (ion counts), *Synechocystis* peptides are detected at all intensities of ΣPep, and the errors in estimating the fraction of *Synechocystis* peptides are symmetrical; however, the *magnitude* of the error is a strong function of the ΣPep signal (Figure S12, Supporting Information). The consequence is numerous cases of false detection of 100% of a single endmember in low-abundance

dx.doi.org/10.1021/ac502494b | Anal. Chem. 2014, 86, 8514−8520

wells. These wells actually contain multiple components, sometimes from both species, that are below the QTOF-MS/MS detection limits. Ordinary least-squares (OLS) regression then yields significant error in predicting the intercepts for δ_{Allo} and δ_{Syn} (Figures 6 and S12, Supporting Information). The

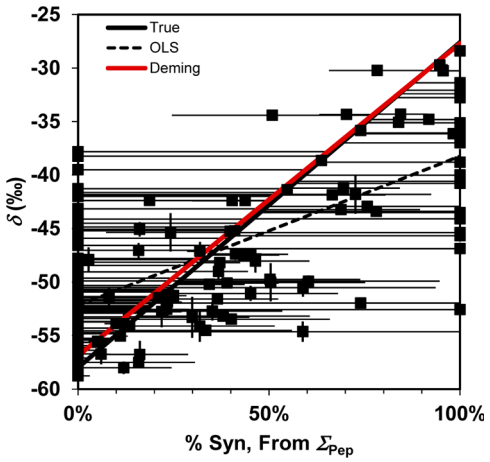


Figure 6. All values of δ'_m plotted against the % *Synechocystis* peptide signal ($\Sigma Pep_{Syn}/\Sigma Pep_{Total}$). Error in δ'_m ($\sigma_{\delta m}$) in most cases is smaller than the size of the symbol. Large errors in the % *Synechocystis* estimates are a strong function of the concentration (ΣPep ion counts). Deming regression accurately predicts the true endmembers, while OLS regression does not. See also Figure S12, Supporting Information.

root-mean-square error (RMSE) of the OLS line in Figure 6 and the average error of the raw data (Figure S12, Supporting Information) both are ±20%. The predicted endmember values for δ_{Allo} and δ_{Syn} would then be −52.2‰ and −38.2‰, i.e., wrong by 6−10‰.

Such a phenomenon is known as regression dilution, or attenuation of the regression slope, and it can be corrected by an errors-in-variables model that accounts for errors in x as well as in y. It is particularly critical in cases such as that shown here, in which the relative standard deviation (σ/signal) is ≫ for the independent variable ΣPep (x) than it is for the dependent variable δ'_m (y). We adopt the approach of Deming regression, which assumes a constant ratio of variances, $\Delta = \sigma_{\delta m}^2/\sigma_{\Sigma Pep}^2$ (using $\sigma_{\delta m} = \pm$ 0.5‰; $\sigma_{\Sigma Pep}$ = 20%; further details in the Supporting Information). Fitting the data in Figure 6 with the Deming regression corrects the slope bias and yields predicted values for δ_{Allo} and δ_{Syn} within 1σ of the true mean. These endmember predictions are shown as gray dashed lines in Figure 5.

DISCUSSION

Although our proof-of-concept trial contains only two species, the data can be used to simulate the performance of the method for more complex systems. The questions of interest are (i) what is the minimum number of data points required to obtain an accurate estimate of δ for a taxon, and (ii) how many isotopic bins (possibly representing functionally different taxa) can be detected within a sample?

To answer question (i), we ran a resampling simulation, selecting random subsets of 100, 75, 50, 25, and 10 data points from the 154 original data points (n = 100 bootstrap replicates of each). The Deming regression was repeated for each simulation, and endmember predictions for δ_{Allo} and δ_{Syn} were tabulated (Table S6, Supporting Information). None of the means for these trials was statistically different from the true means for δ_{Allo} and δ_{Syn} (t-test, p = 0.05), but the smaller the number of data points, the worse the standard deviation of the estimates. Trials with ≥25 data points yielded predictions for δ_{Allo} with $\sigma < \pm$ 2‰, while trials with ≥75 data points were required to yield predictions for δ_{Syn} with $\sigma < \pm$ 2‰. The likely reason for this unequal performance is the higher apparent numbers and concentrations of *A. vinosum* proteins in the overall sample (>2:1 ratio of identifications). For example, in a simulation of 25 data points, *Synechocystis* is likely to be represented only 7 times, in contrast to 18 protein identifications for *A. vinosum*. This suggests that the critical unit is not the absolute number of δ values measured but rather that number times the fractional abundance of the taxon. The results here suggest that n = 15−20 protein identifications are required to yield a value of δ that is within ±2‰ of "true" or n = 30−40 to yield a value of δ within ±1‰ of "true" (Table S6, Supporting Information).

This calculation also provides a partial answer to question (ii). Across a P-SIF sample of 200 measurements that also has evenly distributed abundances of taxa (not necessarily a reasonable assumption for a natural system), we could determine values of δ for 5−10 taxonomic bins. This may be sufficient to describe the dominant members of the community, but in the absence of additional improvements to our detection limits, it will not characterize the minor members. In addition, a more complete answer to (ii) also requires assessment of at least two other factors. The functional resolution will depend on the range of values of δ detected in the total system [here, −28‰ − (−58‰) = 30‰] relative to the standard deviation of a single taxon (1σ = ± 1‰); i.e., in the present example, there would have been a maximum of 30 potential isotopic niches. The functional resolution also will depend on the user-defined definition of a taxonomic bin; e.g., will all sequences detected from Cyanobacteria be counted within a single taxonomic bin or will they be subdivided?

Finally, there is additional information contained within the δ values that is entirely independent of the peptide sequence data. Namely, the apparently tight range of carbon isotopes within individual species indicates that the distribution of δ values within a total, mixed population will be informative. It may be possible to make an isotopic definition of the minimum functional diversity of heterogeneous microbial communities based solely on the breadth and shape of the statistical distribution of all measurements of δ. This approach would be analogous to the isotopic definition of trophic structure within macroscopic communities, based on the enrichment of ^{13}C with successive trophic levels.[12] It also should be possible to explore carbon substrate utilization within cocultures and syntrophic systems, especially in obligate symbiotic associations. P-SIF would then be a complement to single-cell techniques, including SIMS.

CONCLUSIONS

The relatively rapid throughput of P-SIF is useful for ecosystem research and other studies that benefit from assessing population-wide patterns in stable carbon isotope distribution. Although here we focus on natural isotope distributions, our method does not preclude a hybrid approach with SIP. Pulse-chase labeling could increase the effective range of the $\delta^{13}C$ signal, enabling detection of a broader range of functionally distinct taxa. Experiments using physiological levels of

dx.doi.org/10.1021/ac502494b | *Anal. Chem.* 2014, 86, 8514−8520

substrates that include ^{13}C labels likely would succeed and may also be useful to biomedical research.

■ ASSOCIATED CONTENT

Supporting Information

Additional information and figures as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pearson@eps.harvard.edu. Tel: 617-384-8392.

Present Addresses

[†]W.M.: Max Planck Institute for Marine Microbiology, Biogeochemistry, Celsiusstrasse 1, 28359 Bremen, Germany.
[‖]R.J.B.: Schlumberger-Doll Research Center, 1 Hampshire St., Cambridge, Massachusetts 02139.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Gordon and Betty Moore Foundation, the NSF-Dimensions of Biodiversity program, a Marie Curie International Outgoing Fellowship within the seventh European Community Framework Programme, and Harvard University for their generous financial support. We are enormously grateful to the following people for their advice and assistance, without whom none of this would have been possible: John Hayes, Alex Sessions, Suni Shah, Jill Mikucki, Andreas Hilkert, and Tony Fernandez were critical to the development of our SWiM-IRMS + autosampler system; Mak Saito, Dawn Moran, Patricia Clark, and Lisa Lapidus provided invaluable advice on proteomics and protein handling; Dan Rogers and Peter Girguis assisted with the *A. vinosum* culture; Brendan Meade and Steve Beaupre provided advice on quantitative methods; and finally we thank Susan Carter, who manages our laboratory and maintains the instrumentation.

■ REFERENCES

(1) Stahl, D. A.; Hullar, M.; Davidson, S. In *Prokaryotes. A Handbook on the Biology of Bacteria: Symbiotic Associations, Biotechnology, Applied Microbiology*, 3rd ed.; Springer: New York, 2006; Vol. *1*, pp 299−327.
(2) Turnbaugh, P. J.; Ley, R. E.; Mahowald, M. A.; Magrini, V.; Mardis, E. R.; Gordon, J. I. *Nature* **2006**, *444*, 1027−1031.
(3) Reed, H. E.; Martiny, J. B. H. *FEMS Microbiol. Ecol.* **2007**, *62*, 161−170.
(4) Orphan, V. J.; House, C. H.; Hinrichs, K. U.; McKeegan, K. D.; DeLong, E. F. *Science* **2001**, *293*, 484−487.
(5) Thiel, V.; Heim, C.; Arp, G.; Hahmann, U.; Sjovall, P.; Lausmaa, J. *Geobiology* **2007**, *5*, 413−421.
(6) Dekas, A. E.; Poretsky, R. S.; Orphan, V. J. *Science* **2009**, *326*, 422−426.
(7) Jehmlich, N.; Schmidt, F.; Taubert, M.; Seifert, J.; von Bergen, M.; Richnow, H. H.; Vogt, C. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 1871−1878.
(8) Jehmlich, N.; Schmidt, F.; Taubert, M.; Seifert, J.; Bastida, F.; von Bergen, M.; Richnow, H. H.; Vogt, C. *Nat. Protoc.* **2010**, *5*, 1957−1966.
(9) Radajewski, S.; Ineson, P.; Parekh, N. R.; Murrell, J. C. *Nature* **2000**, *403*, 646−649.
(10) Mayali, X.; Weber, P. K.; Brodie, E. L.; Mabery, S.; Hoeprich, P. D.; Pett-Ridge, J. *ISME J.* **2012**, *6*, 1210−1221.
(11) Hayes, J. M. *Rev. Mineral. Geochem.* **2001**, *43*, 225−277.
(12) Deniro, M. J.; Epstein, S. *Geochim. Cosmochim. Acta* **1978**, *42*, 495−506.
(13) Blair, N.; Leu, A.; Munoz, E.; Olsen, J.; Kwong, E.; Desmarais, D. *Appl. Environ. Microbiol.* **1985**, *50*, 996−1001. Hayes, J. M. *Mar. Geol.* **1993**, *113*, 111−125.
(14) Pepaj, M.; Holm, A.; Fleckenstein, B.; Lundanes, E.; Greibrokk, T. *J. Sep. Sci.* **2006**, *29*, 519−528.
(15) Stobaugh, J. T.; Fague, K. M.; Jorgenson, J. W. *J. Proteome Res.* **2013**, *12*, 626−636.
(16) Sessions, A. L.; Sylva, S. P.; Hayes, J. M. *Anal. Chem.* **2005**, *77*, 6519−6527.
(17) Hayes, J. M.; Freeman, K. H.; Popp, B. N.; Hoham, C. H. *Org. Geochem.* **1990**, *16*, 1115−1128.
(18) Merritt, D. A.; Freeman, K. H.; Ricci, M. P.; Studley, S. A.; Hayes, J. M. *Anal. Chem.* **1995**, *67*, 2461−2473.
(19) Matthews, D. E.; Hayes, J. M. *Anal. Chem.* **1978**, *50*, 1465−1473.
(20) Caimi, R. J.; Brenna, J. T. *Anal. Chem.* **1993**, *65*, 3497−3500.
(21) Brand, W. A.; Dobberstein, P. *Isot. Environ. Health Stud.* **1996**, *32*, 275−283.
(22) Thomas, A. T.; Ognibene, T.; Daley, P.; Turteltaub, K.; Radousky, H.; Bench, G. *Anal. Chem.* **2011**, *83*, 9413−9417.
(23) Nelson, D. M.; Hu, F. S.; Mikucki, J. A.; Tian, J.; Pearson, A. *Geochim. Cosmochim. Acta* **2007**, *71*, 4005−4014.
(24) MacGregor, B. J.; Bruchert, V.; Fleischer, S.; Amann, R. *Environ. Microbiol.* **2002**, *4*, 451−464.
(25) Pearson, A.; Sessions, A. L.; Edwards, K. J.; Hayes, J. M. *Mar. Chem.* **2004**, *92*, 295−306.
(26) Tsao, L. E.; Robinson, R. S.; Higgins, M. B.; Pearson, A. *Org. Geochem.* **2012**, *49*, 96−99.
(27) O'Farrell, P. H. *J. Biol. Chem.* **1975**, *250*, 4007−4021.
(28) Görg, A.; Weiss, W.; Dunn, M. J. *Proteomics* **2004**, *4*, 3665−3685.
(29) Paspaltsis, I.; Berberidou, C.; Poulios, I.; Sklaviadis, T. *J. Hosp. Infect.* **2009**, *71*, 149−156.
(30) Macko, S. A.; Fogel, M. L.; Hare, P. E.; Hoering, T. C. *Chem. Geol.* **1987**, *65*, 79−92.