Quantitative tracking of isotope flows in proteomes of microbial communities

Running title: Proteomic stable isotope probing

Chongle Pan\textsuperscript{1,2,3,#,*}, Curt R. Fischer\textsuperscript{4,*}, Doug Hyatt\textsuperscript{3}, Benjamin P. Bowen\textsuperscript{5}, Robert L. Hettich\textsuperscript{2}, Jillian F. Banfield\textsuperscript{4,+}

\textsuperscript{1} Computer Science and Mathematics Division, Oak Ridge National Laboratory
\textsuperscript{2} Chemical Science Division, Oak Ridge National Laboratory
\textsuperscript{3} BioScience Division, Oak Ridge National Laboratory
\textsuperscript{4} Dept. of Earth and Planetary Science, University of California, Berkeley
\textsuperscript{5} Life Sciences Division, Lawrence Berkeley National Laboratory

\textsuperscript{#} Contact C.P. (panc@ornl.gov) for questions on proteomic methodology and informatics.

\textsuperscript{+} Contact J.F.B. (jbanfield@berkeley.edu) for questions on microbial community dynamics.

\textsuperscript{*} Both authors contributed equally to this work.

Keywords: Stable isotope probing/proteomics/microbial communities.
Abstract:

Stable isotope probing (SIP) has been used to track nutrient flows in microbial communities, but existing protein-based SIP methods capable of quantifying the degree of label incorporation into peptides and proteins have been demonstrated only by targeting usually less than 100 proteins per sample. Our method automatically (i) identifies the sequence of and (ii) quantifies the degree of heavy atom enrichment for thousands of proteins from microbial community proteome samples. These features make our method suitable for comparing isotopic differences between closely related protein sequences, and for detecting labeling patterns in low-abundance proteins or proteins derived from rare community members. The proteomic stable isotope probing (SIP) method was validated using proteome samples of known stable isotope incorporation levels at 0.4%, ~50%, and ~98%. The method was then used to monitor incorporation of $^{15}$N into established and regrowing microbial biofilms. The results indicate organism-specific migration patterns from established into regrowing communities and provides insight into metabolism during biofilm formation. The SIP-proteomics method can be extended to many systems to track fluxes of $^{13}$C or $^{15}$N in microbial communities.

Introduction:

Nitrogen or carbon flows from $^{15}$N- or $^{13}$C-enriched substrates into the biomass of a microbial community can be traced by measuring the incorporation of these stable isotopes into biomarkers such as lipids, nucleic acids, and proteins. This technique, called stable isotope probing (SIP), has been used to elucidate metabolic activities in microbial communities [1]. The first example of SIP used lipids as the biomarker [2]. A low level of $^{13}$C incorporation in lipids can be
determined by gas chromatography-mass spectrometry, but it is generally difficult to link lipids to specific microorganisms. More commonly, DNAs or RNAs are used as the biomarkers for SIP analysis [3]. Isotopically labeled nucleic acids are separated from unlabeled nucleic acids using buoyant density gradient centrifugation. Analysis of the separated nucleic acids provides direct information about which microorganisms have incorporated the label. However, gradient centrifugation can only resolve nucleic acids with large differences in the degree of label incorporation. Recently a protein-based SIP method was developed which uses mass spectrometry (MS) to determine $^{13}\text{C}$ incorporation extents of peptides and proteins [4-9]. Because the degree of label incorporation in proteins and peptides can be determined to high resolution by mass spectrometry, and as proteins and peptides contain sequence information that links each molecule to its organism of origin, protein-based SIP methods enable determination of low levels of isotope incorporation into microorganisms that can be resolved at the strain level.

Early protein SIP studies determined the $^{13}\text{C}$ atom% of up to 38 proteins from a single organism in a pure culture or an enrichment culture. Organisms studied included *Methylibium petroleiphilum* [8] and *Aromatoleum aromaticum* [4]. In both studies, proteins of unknown $^{13}\text{C}$ atom% were identified by matching their gel spots from 2D gel electrophoresis to spots from unlabeled proteomes. The $^{13}\text{C}$ atom% was estimated using least-square fitting analysis of isotopic distributions of detected peptides. The measured $^{13}\text{C}$ atom% of a limited number of proteins allowed tracking of $^{13}\text{C}$ from a substrate into a species of interest, but this number of proteins is insufficient for intra-organismal, pathway-resolved comparisons of label incorporation dynamics. More recently, a new method based on peptide decimal place slope was developed for protein SIP [7, 8]. This method requires aggregation of at least 100 peptides for
precise $^{13}$C atom% estimation and allows estimation of aggregate $^{13}$C atom% in an organism’s proteome. However, it cannot resolve individual proteins’ atom% or pathway-specific differences.

Here, we report a proteomic SIP method that can determine $^{15}$N atom% of thousands of identified proteins from multiple strains and species in a model microbial community. Compared to the existing SIP methods, this new proteomic SIP method provides a deep coverage of strain-resolved datasets with variable extents of incorporation of $^{15}$N/$^{13}$C. The approach provided new insights into the metabolic activity by allowing metabolic comparison of established and regrowing multispecies biofilm communities.

**Results:**

*Validation of the proteomic SIP method using proteome samples of known $^{15}$N atom%.*

The proteomic SIP method was validated using laboratory grown microbial communities derived from an acid mine drainage (AMD) system previously extensively studied by proteomics [10-13]. In the natural system, these extremophilic communities grow on the surface of highly acidic and metal-rich solutions (pH ~0.8 and near-molar [Fe$^{2+}$]) and harvest energy from pyrite oxidization. Despite the harsh conditions, the community has a primary productivity of ~0.5 g C m$^{-2}$day$^{-1}$, comparable to many terrestrial and aquatic ecosystems [11]. The AMD community has been established as a model microbial community for metagenomic [14], community proteomic, and ecological studies [13].
In the current study, three AMD biofilms were grown in laboratory bioreactors in which nitrogen was solely supplied as ammonium with either 0.4 atom% $^{15}\text{N}$ (natural abundance), $\sim$50 atom% $^{15}\text{N}$, or $\sim$98 atom% $^{15}\text{N}$. For the latter two conditions, uncertainty in the $^{15}\text{N}$ atom percentage of ammonium supplied in the medium stems mostly in uncertainty in the isotopic purity of enriched ($^{15}\text{NH}_4)_2\text{SO}_4$) purchased from our supplier, which is rated as “$>98$” atom% $^{15}\text{N}$. (Medium with $\sim$50 atom% $^{15}\text{N}$ was prepared by weighing labeled and unlabeled ammonium sulfate on an analytical balance and is precise to three decimal places.) Biofilms were grown over a 3-week period from low concentrations of planktonic cells with constant medium composition. All proteins in each biofilm should have approximately the same $15\text{N}$ atom% as the $15\text{N}$ atom% of the ammonium used in the growth medium as the nitrogen source.

Proteomes were digested by trypsin and analyzed using 24-hour 2-dimensional liquid chromatography tandem mass spectrometry (2D-LC MS/MS) [15]. In these validation experiments, we used both an LTQ Orbitrap XL instrument [16] and an LTQ Orbitrap Velos instrument [17] (Thermo Fisher Scientific) to investigate the effects of instrument types on the number and false discovery rate of protein identifications. High-resolution MS/MS scans were obtained from both instruments (Figure 1A). The isolation window was set to be $\pm 2.5$ Da wide to include most of the high-intensity isotopic peaks of peptides at varying $^{15}\text{N}$ atom%. A sample set of acquired MS/MS spectra are provided in Supplemental Figure S1. A free open-source algorithm, Sipros, was developed to search the MS/MS data against a protein sequence database at a range of possible isotopic enrichment percentages. The algorithm was released at http://code.google.com/p/sipros/. Sipros built a theoretical spectrum for every tryptic peptide
predicted from protein sequences at every $^{15}$N atom% from 0% to 100% at 1% intervals (Figure 1B). Experimental spectra were searched against all these theoretical spectra to identify both peptides’ sequences and their $^{15}$N atom%. On average it consumed approximately 2000 processor hours (i.e. ~20 hours wall-clock time using 100 AMD Opteron processors) to process a 24-hour 2D-LC MS/MS dataset.

The $^{15}$N atom% values of proteins were estimated from the $^{15}$N atom% of their constituent peptides. In a SIP experiment, a protein may have multiple isotopologues at distinct $^{15}$N atom%. Peptides with similar $^{15}$N atom% were grouped together using a hierarchical clustering algorithm (Figure 1C). A cluster containing more than two peptides of similar $^{15}$N atom% was considered representing an isotopologue of the protein. The unsupervised clustering method allowed automated identification of an unknown number of distinct $^{15}$N isotopologues of a protein in a proteome. To remove redundant protein identifications that share non-unique peptides, protein identifications with identical sets of peptides were grouped together and proteins whose peptide sets are strict subsets of another protein’s peptide set were removed. The remaining identifications were called non-redundant protein identifications. Finally, if a protein has two distinct $^{15}$N isotopologues, the abundance ratios for isotopologues pairs were estimated from selected ion chromatograms of their constituent peptides using the ProRata program [18, 19].

Overall, many non-redundant proteins were identified at low false discovery rates (< 2%) from six members of the AMD community in all three validation samples (Figure 2 and Supplemental Table 1). Estimated $^{15}$N atom% of proteins agreed well with the experimentally imposed $^{15}$N abundances (Figure 2A and 2B). Because the LTQ Orbitrap Velos scanned almost twice as fast
as LTQ Orbitrap XL, many more proteins were identified with this instrument, and it was used exclusively for future experiments. However, the data in Supplemental Table 1 shows that the Sipros technique is still effective on an LTQ Orbitrap XL. Among the three validation samples, the 50% labeled sample had the lowest numbers of identified proteins, highest false discovery rates, and lowest \(^{15}\text{N} \) atom% precision (Figure 2C). This was expected because peptides of 50\(^{15}\text{N} \) atom% have more complex isotopic packets than packets for peptides at other \(^{15}\text{N} \) atom%. In summary, this analysis showed that the new method can confidently identify proteins of varying \(^{15}\text{N} \) atom% and accurately estimate their \(^{15}\text{N} \) atom%.

**Proteomic SIP during recolonization and regrowth of laboratory AMD communities.**

*Labeling patterns reveal cell migration.* AMD biofilms grow on the surface of streams and pools and are periodically removed by floods or other natural events. The community then recolonizes the cleared space and develops into thick mature biofilms [10, 20]. The proteomic SIP method was used to study the development of the AMD community to provide biological insight into the colonization process. This analysis used multispecies biofilms grown in laboratory bioreactors. Similar communities were shown previously to closely replicate natural communities, both in terms of membership and activity [11].

In the experiment, we grew the community in laboratory bioreactors supplied with natural abundance \(^{15}\text{N} \), then removed a piece of unlabeled biofilm and simultaneously switched to a medium containing ~98\(^{15}\text{N} \) (Supplemental Figure S2). 105 hours later we sampled both the original established (sample A1) and nascent regrowing (B1) biofilms. This procedure was
repeated 274 hours later (yielding sample A2 from the original biofilm, and sample B2 from the second regrowth of the biofilm). From 1814 to 2407 non-redundant proteins were identified from each of these samples using LTQ Orbitrap Velos (Supplemental Table 2). The most highly represented organism in all samples was *Leptospirillum* group III, for which 1200 proteins were detected in one or more of the four samples (~45% of the predicted proteins) [21]. Figure 3B shows the microorganism-resolved distributions of the unlabeled (<5% ¹⁵N atom %) and labeled proteins (>80% ¹⁵N atom %) in the four samples. Proteins uniquely assignable to ten different Bacteria and Archaea and two virus populations were identified.

Figure 3A shows the extent of ¹⁵N atom% labeling in proteins across the proteomes. Most proteins in all four samples were either unlabeled (<5% ¹⁵N atom %) or highly labeled (>80% ¹⁵N atom %). Most proteins in the established biofilm samples remained unlabeled after the medium switch. In the regrowing biofilm samples, there were both labeled and unlabeled proteins.

It was intriguing that biofilm regrowing in a ~98% ¹⁵N medium contained significant numbers and amounts of unlabeled proteins (Figure 3). Three distinct scenarios could explain the presence of unlabeled proteins in the regrowing biofilm. First, the unlabeled proteins could have originated by carryover contamination, i.e. inadvertent inclusion of some established biofilm in the samples of regrowing biofilm. The data in Figure 3B discount this scenario, as unlabeled proteins in regrowing biofilm (sample B1) had organismal distributions distinct from unlabeled proteins in established biofilm (sample A1). An alternate scenario is that small-molecule sources of organic nitrogen were trafficked between the established and regrowing biofilms, *i.e.* spatially...
separated isotopic recycling. The $^{14}\text{N}$ would mix with $^{15}\text{N}$ from the medium at the level of \textit{de novo} amino acid biosynthesis, and the mixed nitrogen pool would result in the biosynthesis of proteins with intermediate labeling percentages. The data in Figure 3A discounts this scenario; only highly unlabeled or highly labeled proteins were detected. The final scenario, consistent with Figure 3, is that whole cells (and their unlabeled proteomes) migrated into the regrowing biofilm. The distinct organismal distributions of unlabeled proteins in regrowing vs. established biofilms is indicative of selective migration of certain species into regrowing biofilm. We cannot exclude the possibility subsequently to or in tandem with cell migration, isotope recycling processes such as protein turnover also contribute to the observed data, but as shown in Supplemental Figure S2, near the time of $^{15}\text{N}$ addition to the bioreactor, biofilm was completely excised from the region in which regrowing biofilm subsequently appeared.

Many proteins were identified as two distinct $^{15}\text{N}$ isotopologues. The abundance ratios of the unlabeled and labeled isotopologues were estimated (Supplemental Table 3). Figure 3C shows the distributions of protein abundance ratios by microorganisms. Unlabeled isotopologues dominated in the established biofilm samples (Figures 3A and 3C), demonstrating very slow rates of $^{15}\text{NH}_4^+$ incorporation into proteins. This finding demonstrates, for the first time, low \textit{de novo} protein synthesis from $\text{NH}_4^+$ in these mature communities. In the regrowing biofilm samples, the labeled isotopologues of proteins were more abundant than the unlabeled isotopologues. The detection of labeled proteins indicated significant uptake of $^{15}\text{NH}_4^+$ for new protein biosynthesis.
Labeling patterns of dominant community members. As shown in Figure 4, the majority of labeled proteins in the regrowing biofilm derive from *Leptospirillum* Group III, especially in Sample B1. 234 *Leptospirillum* Group III proteins were identified only in the early regrowth sample (B1), and all of these were completely $^{15}$N labeled (see proteins highlighted by a red bar in the first panel of Figure 4). We attribute the identification of these proteins to a combination of the high abundance of this organism, which enabled deep proteome sampling, and broad metabolic activity associated with extensive cell division and growth. Of these 234 proteins, 28% (66 proteins) had no annotated function, but proteins with no annotated function comprised just 12% of *Leptospirillum* group III proteins identified in more than just sample B1. This difference is statistically significant (Fisher’s two-tailed exact test, $p = 2.6 \times 10^{-8}$), and may indicate that biological processes required for early biofilm colonization are largely uncharacterized.

We mapped protein isotope composition information onto biochemical pathways to seek evidence for either rapid turnover of specific proteins or up-regulation of entire pathways. The analysis results are shown in Supplemental Appendix 1. In some pathways, such as the reductive TCA cycle (Figure 5A), isotopologue abundance varied widely among member proteins, indicating $^{15}$N incorporation levels were determined primarily by protein-specific turnover rates. However, for other pathways, labeling patterns were broadly consistent among member proteins. Examples include large ribosomal subunit proteins (Figure 5B), the pyruvate metabolic pathway and chemotaxis and heme and cobalamin biosynthesis during early regrowth (Sample B1, as shown in Supplemental Appendix 1). This suggests that these specific functionalities were up-regulated in the early regrowth biofilm (sample B1).
Surprisingly, the later regrowth biofilm sample (B2) contained many unlabeled *Leptospirillum* Group III proteins. As noted above, this finding is attributed to immigration from the pre-existing biofilm. Proteins in sample B2 which are predominantly unlabeled must be expressed in migrating cells and must also not be highly expressed in cells involved in *de novo* biosynthesis and growth in the re-growing biofilm. For example, in *Leptospirillum* group III, aminoacyl tRNA synthetases were significantly overrepresented in unlabeled proteins in B2 as compared to labeled proteins in B2 (Fisher two-sided exact test, $p = 2.8 \times 10^{-6}$), indicating that large amounts of these proteins were carried into the sample by *Leptospirillum* Group III cells migrating from already-established, $^{14}$N-rich regions of the biofilm. A large role for late immigration could not have been predicted based without dynamic labeling experiments, which are rarely conducted in studies of complex, multispecies microbial communities. Notably, despite evidence for migration, we do not identify both isotopologues for many abundant proteins. There are *Leptospirillum* Group III proteins in the late regrowth (B2) biofilm that are predominantly unlabeled for which strongly labeled versions were identified in the early regrowth stage. We suggest that these may be slow turnover proteins that are much more important to *Leptospirillum* living in mature biofilms and the high abundance of unlabeled isotopologues masks the signal from newly biosynthesized ones.

In contrast to Samples B1 and B2 of the regrowing biofilm, samples A1 and A2 from established biofilm had low numbers of $^{15}$N-enriched isotopologues and low abundance ratios of $^{15}$N isotopologues to natural isotopologues. These observations indicate that most proteins had not turned over on the time scale of the experiment, more than 300 hours. However, a small subset of labeled proteins were detected in established biofilms, which (i) demonstrates that de novo
amino acid biosynthesis was occurring in established biofilms, and (ii) suggests that these labeled proteins have unusually short half-lives. In the UBA and 5wayCG *Leptospirillum* Group II proteomes, a predicted cytochrome (UBA_LeptoII_Scaffold_8062_GENE_147 # Locus_tag: UBAL2_80620147 # Function: Probable cytochrome c, class I) was the sole protein for which the $^{15}$N-rich isotopologue was more abundant than its $^{14}$N-labeled counterpart in both A1 and A2. In *Leptospirillum* Group III, 7 proteins of various functions and activities met this criterion. These proteins share no common function. We infer that their high level of labeling indicates rapid turnover due to exposure to the extracellular environment.

Notably, the Archaeal G-, A-, and E-plasma, *Ferroplasma* I and *Ferroplasma* II proteins were mostly unlabeled in samples A1, A2, and B2, and very few Archaeal proteins were identified in sample B1. The dearth of unlabeled archaeal proteins in sample B1 and their presence in sample B2 (Figure 3B) suggests that Archaea migrate into developmentally mature biofilms. Previous studies found a higher archaeal species abundance in more mature compared to newly formed biofilms [13, 20, 22]. The protein SIP results indicate this may be primarily due to immigration rather than *in situ* cell division following migration of a few colonists. Given results documenting both early and late Bacterial migration, the results also demonstrate that migration preferences of species changes as biofilms developed.

*Labeling patterns of rare community members.* Proteins encoded by viruses and unclassified genome fragments and other non-unique proteins were categorized as “Others” in Figure 3B. It is interesting that this group includes 12 proteins from the bacteriophage AMDV1. This phage is known to replicate in *Leptospirillum* but has not been detected as a prophage. In both established
biofilms samples, all phage proteins were unlabeled whereas in sample B1 the viral proteins were either almost completely labeled or occurred exclusively as the labeled isotopologue. In sample B2, we detected a mixture of unlabeled and essentially completely labeled AMDV1 proteins. In contrast, the single protein from an Archaeal virus that was detected in A2 and B2 occurred exclusively in the unlabeled form. The observed labeling patterns are consistent with the labeling patterns of the hosts in which the viruses replicate (Fig. 3B). Thus, the labeling results indicate that detected viral proteins were synthesized in Bacterial and Archaeal host cells close to the time of sampling.

Actinobacteria are low-abundance members of acid mine drainage consortia from the Richmond Mine. We detected between 4 and 12 actinobacterial proteins in the four samples A1, A2, B1, and B2. In B2, both highly labeled and unlabeled isotopologues of an actinobacterial GroEL protein (UNLA1_11998005 # Function: COG0459 Chaperonin GroEL) were detected. Quantification by ProRata revealed that the $^{15}$N-labeled isotopologue was 9 times more abundant than the unlabeled isotopologue, showing that actinobacteria were actively metabolizing nitrogen and/or synthesizing protein during our experiment. Several other $^{15}$N-labeled isotopologues of Actinobacterial proteins were also detected in samples B1 and B2, supporting this finding.

Nanoarchaea of the ARMAN lineage are recently discovered, low-abundance members of the AMD community [23]. Detected ARMAN proteins were exclusively in the unlabeled state. Only 7 of the 33 proteins identified had no assigned function and those with functional annotations were largely associated with metabolism, growth, or defense mechanisms. ARMAN lineages are highly divergent relative to other Archaea and little is known about their physiology,
Despite the availability of three near-complete genome sequences [24]. Some genome features suggest that they are heterotrophs. Three dimensional cryogenic electron tomographic imaging has documented intimate cytoplasmic connections between the cells and coexisting Thermoplasmatales lineage Archaea [24], although the nature of the association is unknown. The finding that ARMAN nanoarchaea did not take up ammonium during the experiment indicates that either the cells are inactive at the time scales of our experiment or that they derive their nitrogen from unlabeled sources, probably coexisting Archaea. Importantly, however, ARMAN proteins detected in the $^{15}$N-enriched validation samples (Figure 2) were labeled at enrichment ratios of 95.5 atom% (ARMAN proteins in the >98% $^{15}$N validation sample) and between 10 and 49 atom% (50 atom% $^{15}$N validation sample), showing that ARMAN can be metabolically active in laboratory-grown whole-community biofilms.

**Discussion:**

*Comparison of proteomic SIP to the existing SIP methods*

The proteomic SIP method is enabled by a novel approach for peptide identification and $^{15}$N atom% estimation from high-resolution MS/MS spectra using the Sipros algorithm. Conventionally, the extent of stable isotope incorporation in a known peptide can be estimated by comparing its measured isotopic distribution to a series of expected isotopic distributions calculated at different atom% of a stable isotope (Yergey 1983). This has been used for protein turnover ratio analysis (Cargile et al. 2004), protein assay prior to NMR analysis (Choudhary et al. 2006), and protein-based SIP (Jehmlich et al. 2008). More recently, a new method based on decimal place slope was developed for $^{13}$C atom% estimation. Because the difference between
$^{12}\text{C}$ and $^{13}\text{C}$ is 1.003355 Da, a peptide with higher $^{13}\text{C}$ atom\% has a higher slope for the digits in
the third and fourth place after the decimal point. Because it is challenging to achieve such an
accurate mass measurement on every peptide, at least 100 peptides were required for accurate
$^{13}\text{C}$ atom\% estimation. Therefore, unlike the conventional method or our new method, the
decimal place slope method cannot provide accurate $^{13}\text{C}$ atom\% estimation for individual
proteins.

Our method based on the Sipros algorithm offers two advantages over the existing methods.
First, existing methods requires an offline procedure to identify the sequence of peptides with
variable label incorporation. For example, the original protein-SIP method requires a parallel 2D
gel electrophoresis of an unlabeled proteome with the SIP proteome sample for protein
identification. This limits the throughput of the analysis. Our method performs peptide
identification and $^{13}\text{C}/^{15}\text{N}$ atom\% estimation in a single measurement in a fully automated
manner. Second, to estimate $^{13}\text{C}/^{15}\text{N}$ atom\%, the existing methods use isotopic distributions of
intact peptides in MS1 scans, whereas Sipros uses isotopic distribution of peptides’ fragments in
MS2 scans. Because many peptides that are barely detectable in MS1 scans can yield high-
quality MS2 scans, Sipros should be able to provide accurate estimations of $^{15}\text{N}$ atom\% for many
low-abundance peptides. Since the $^{13}\text{C}/^{15}\text{N}$ atom\% of a peptide is estimated based on isotopic
distributions of multiple fragment ions rather than a single isotopic distribution of the parent ion,
the estimation should be more robust to random noise interference. The two advantages of our
method allowed identification thousands of partially labeled proteins and quantification of their
$^{15}\text{N}$ atom\% and relative abundances in a natural microbial community. This represents a
significant methodology improvement from previous studies in which the $^{13}\text{C}$ atom\% of up to
only 38 proteins were determined from a single organism in a pure culture or an enrichment culture [4-6, 9].

While the existing SIP methods focus on tracking isotope flows between microorganisms based on the analysis of a small set of DNAs, RNAs, proteins, or lipids, the proteomic SIP method allows system-level functional analysis of microorganisms involved in the isotope transfer using the large number of identified proteins. In a perspective article on SIP (Dumont and Murrell 2005), Dumont and Murrell envisioned the emergence of genomic SIP and transcriptomic SIP in which $^{13}$C enriched DNAs and RNAs are isolated and sequenced and the biological insights that these potential methods can provide. We believe proteomic SIP is the first method that demonstrated the potential of combining the high-throughput proteomics technology with the SIP technique.

**Application of proteomic SIP to complex communities**

The study demonstrates that the proteomic SIP method is able to identify proteins from Bacteria, Archaea and viruses in natural communities at the strain level and to characterize sub-populations that differ in the extent of their isotopic labeling. We unexpectedly observed labeling patterns indicative of cell migration into regrowing regions of biofilm, illustrating that our method is useful for study of complex biological phenomena. In established biofilms, most proteins had very little $^{15}$N incorporation over the time scale of the experiment. We suspect this indicates very low protein turnover rates in the mature biofilms, but cannot exclude the possibility that proteins are recycled to their constituent amino acids rapidly but *de novo* rates of amino acid biosynthesis from $\text{NH}_4^+$ are near-zero. Future studies over longer time periods and
using additional isotopes, could lead to better assessment of protein turnover rates and further deconvolution of the mixed effects of protein turnover and cell migration that we observe here.

Using our SIP proteomic methods, it is possible to identify and assess the extent of labeling of proteins from rare members of microbial consortia, as was demonstrated for Actinobacterial, ARMAN-derived, and viral proteins. Quantitative labeling information can be extracted even though proteomes are shallowly sampled and peptide coverage of individual proteins is very poor. Thus, the approach can be used to demonstrate activity of rare members and to quantify flows of nitrogen and, by extension, carbon in more complex communities.

The AMD community was studied here as a model system for proof of principle of the high throughput proteomic SIP method. Several prerequisites are required before applying proteomic SIP to other systems. First, whole-community cultivation must be available so that significant levels of heavy isotope incorporation can be achieved in coexisting organisms under biologically relevant conditions. Second, high quality metagenomic databases should be available for the community of interest so that protein sequence databases can be constructed. While protein identification can be achieved using isolate genome sequences, missing strain-specific candidate peptide sequences will lower protein identification rates and potentially introduce biases. With recent technology advances, more and more communities are now amenable to comprehensive proteogenomics analysis [25]. Provided these challenges can be addressed, the proteomic SIP method will be broadly applicable to communities that comprise more member species than found in the AMD community.
A particular strength of the 2D-LC MS/MS-based SIP method is that many proteins from closely related strains and species can be differentiated, so long as SNPs that differentiate the proteins result in mass shifts in identified tryptic peptides. This is in contrast to approaches that rely on differences in gel migration rates that are unlikely to be sensitive to SNPs and strain-level variation. In the AMD community, approximately half of the protein identifications from the *Leptospirillum* Group II species can be uniquely attributed to one of the species’ two sub-strains (0.3% difference in 16S rRNA gene sequences). Therefore, we expect that many identified proteins from a complex community could be uniquely assigned at the species or strain level. As community complexity increases, fewer pathways will be sampled sufficiently for detailed analysis at the strain level and most organisms may only be amenable to species-level analysis of overall activity levels.

Proteomic SIP is enabled by high-performance computing. More complex communities will require more computing time for Sipros data analysis. The computing time increases linearly with the number of MS/MS spectra and the number of proteins in the metagenomic database. In this study, it took ~20 hours using 100 processors in a cluster to search ~10^5 MS/MS spectra in a validation dataset against a protein database comprising 31966 sequences. The data can also be processed using 101 virtual machines instantiated in commercial cloud, which would cost approximately $340 using Amazon Elastic Compute Cloud at the present unit price of $0.17 per hour for high-CPU on-demand instances. The computing time and cost for processing data from a complex community would increase in proportion to the size of the community metagenomic dataset. However, the computing cost may still be a fraction of the experimental cost of a proteomic SIP study.
Materials and Methods:

Cultivation of standard AMD cultures at varying known $^{15}$N atom%.

Mixed species chemoautotrophic acid mine drainage biofilms were cultivated aerobically at ~40° C in laboratory bioreactors essentially as previously described [11], with the following modifications. The unlabeled standard culture was collected from a previously established biofilm grown in unlabeled medium. The ~50% $^{15}$N-enriched standard culture was grown in a medium containing a mixture of 0.0677 g of >98 atom% ($^{15}$NH$_4$)$_2$SO$_4$ (Sigma-Aldrich, St. Louis, MO) and 0.0638 g of unlabelled (NH$_4$)$_2$SO$_4$ as the sole nitrogen source. The ~98% $^{15}$N-enriched standard culture was grown in a medium containing >98 atom% ($^{15}$NH$_4$)$_2$SO$_4$ as the sole nitrogen source. Spent medium (50 mL) containing planktonic cells from previous, isotopically unlabelled bioreactors served as an inoculum. Biofilm developed in the bioreactors and iron(II) was depleted in the growth medium after 10 days. Then the medium supply to the bioreactors were changed to a single-pass configuration, where the medium continuously flows through bioreactors in a single pass from the fresh-medium reservoir to the waste at a flow rate between 0.5 and 1 mL/min. Biofilms were collected after 10 further days of culture.

$^{15}$N stable isotope probing of the AMD community’s development.

An AMD community culture was grown in a bioreactor for several months in unlabeled medium. A piece of the established but unlabeled biofilm was removed and the medium was switched to one containing ~98% $^{15}$N (>98 atom% ($^{15}$NH$_4$)$_2$SO$_4$). The medium source was switched to $^{15}$N labeled medium (>98 atom% ($^{15}$NH$_4$)$_2$SO$_4$). 105 hours later, a piece of the original established biofilm (sample A1) was collected from a previously undisturbed area and a piece of the nascent
regrowing biofilm (B1) was collected from the previously cleared area. 274 hours later, a second piece of the established biofilm (A2) and a second piece of the regrowing biofilm (B2) were sampled.

**Proteome sample preparation.**

Whole-cell lysates were extracted from the biofilm samples as described previously [10]. The proteins were denatured and reduced with 6 M guanidine and 10 mM dithiothreitol (DTT) (Sigma Chemical Co. St. Louis, MO) at 60° C for 1 hour. The samples were then diluted 6-fold with 50 mM Tris, 10 mM CaCl$_2$ (pH 7.6), and sequencing grade trypsin was added at 1:100 (wt:wt). The first digestion was run overnight at 37° C and after adding additional trypsin, the second digestion was run for 5 hrs at 37° C. Finally, the samples were reduced with 20 mM DTT for 1 h at 60° C and desalted using C18 solid-phase extraction (Sep-Pak Plus, Waters, Milford, MA).

**2D-LC-MS/MS measurements.**

Samples were loaded offline onto a 5-cm-long 250-$\mu$m-I.D. strong cation exchange back column (Luna, Phenomenex). The back column was connected to a 15-cm-long 100-$\mu$m-I.D. C18 reverse-phase PicoFrit column (New Objective) and placed in-line with a U3000 quaternary HPLC (Dionex, San Francisco, CA). The two-dimensional LC separation was performed with eleven salt pulses. The first ten salt pulses were followed by a 2-h reverse-phase elusion gradient from 0% solvent B to 50% and the last one by a 3-h gradient from 0% solvent B to 100%. The LC eluent was directly electrosprayed into the mass spectrometer with an ionization voltage of 4 kV. The AMD standard samples were analyzed using an LTQ-Orbitrap XL instrument in the
following parameters: three data-dependent MS/MS scans following every full scan; MS/MS scans acquired in Orbitrap at resolution 7,500 with two-microscan averaging; full scans acquired in Orbitrap at resolution 30,000 with two-microscan averaging; 35% normalized collision energy; ±2.5 Da isolation window; dynamic exclusion enabled with ±3 Da exclusion window. The AMD standard samples were also analyzed using an LTQ-Orbitrap Velos instrument with identical configurations except for five data-dependent MS/MS scans following every full scan. The AMD SIP samples were analyzed using LTQ-Orbitrap Velos.

**Peptide identification and $^{15}$N atom% estimation.**

All LC-MS/MS datasets were converted from the Xcalibur Raw file format to the FT2 flat file format [26] using the Raxport program (Freely available at http://code.google.com/p/raxport/). The Sipros algorithm v1.0 (Freely available at http://code.google.com/p/sipros) was used to identify peptide sequences and quantify their $^{15}$N atom%. The validation samples were searched against the core AMD database containing 15983 protein sequences and their reverse sequences [10] (http://compbio.ornl.gov/biofilm_amd/). False discovery rates of peptide identification were calculated based on the hits to reverse sequences as described [27]. The regrowth samples were searched against an updated and expanded AMD database containing 57001 protein sequences which include Actinobacterial, ARMAN-derived, and viral proteins. The computation was performed on a 512-core Linux cluster (AMD Opteron processor, RAID protected NFS cluster storage and InfiniBand data network). Every LC-MS/MS dataset was searched using 101 parallel threads spawned on a cluster. Each thread searched the database at a given $^{15}$N atom% ranging from 0% to 100% with a 1% interval. Peptides were predicted from protein sequences using the following trypsin cleavage rules: cleavage after residues K and R; up to two missed cleavage
sites in a peptide; and specific tryptic cleavages at both termini. Candidate peptides for an MS/MS spectrum were found by using seven parent mass windows to filter predicted peptides based on their calculated most abundant masses. Let \( P \) be the parent mass of a spectrum and the seven parent mass windows of size \( \pm 0.04 \) Da were centered at \( P, P \pm 1.0087 \) Da, \( P \pm 2.0174 \) Da and \( P \pm 3.0261 \) Da to accommodate the so-called 1-Da ambiguity in correlating parent ion masses with calculated most abundant isotopic masses. A theoretical MS/MS spectrum were then reconstructed for every candidate peptide and matched to the experimental MS/MS spectrum. All residues in a peptide were assumed to have the same \( ^{15}N \) atom\% for calculating expected isotopic distributions of y and b ions. The observed y and b ions were scored by their mass accuracy and goodness of fit between their expected isotopic distributions and their observed ones [26] and the candidate peptide’s score is the sum of its y and b ions’ scores. The output from all \( ^{15}N \) atom\% searches was aggregated and peptide identifications were filtered using a score cutoff of 25.

Protein identification and \( ^{15}N \) atom\% estimation.

Identified peptides were assigned to proteins based on their amino acid sequences. Peptides belonging to a protein were then clustered by their \( ^{15}N \) atom\% into isotopologue groups using an agglomerative clustering algorithm. The clustering started by putting every peptide in a separate cluster and iterated between merging the two closest clusters into a new cluster and calculating new distances between every pair of clusters. The distance between two clusters was defined as the difference between the average peptide \( ^{15}N \) atom\% of the two clusters. The iteration continued until the distance between the closest clusters was above 20\%. Clusters containing only one peptide were discarded as outliers due to the high false discovery rate for proteins.
identified based on a single peptide. The standard deviation of peptide $^{15}$N atom% in every cluster was calculated and clusters with a standard deviation greater than 10% were filtered out. Remaining clusters were considered representing separate isotopologues of the protein at different protein $^{15}$N atom%. To identify redundant protein identifications, the peptide sets of every two proteins were compared. Protein isotopologues with identical sets of peptides were grouped together and proteins whose peptide sets are strict subsets of other proteins’ peptide sets were removed. The resultant lists of proteins were referred to as non-redundant protein identifications.

**Protein quantification using ProRata.**

For proteins identified at multiple $^{15}$N atom%, the abundance ratio between two isotopologues of a protein was estimated using the ProRata program that was originally developed for quantitative proteomics based on stable isotope labeling [18, 19]. The ProRata program is freely available at http://code.google.com/p/prorata/. While in quantitative proteomics samples proteins generally have two uniform $^{15}$N atom% (a natural $^{15}$N atom% and a known high $^{15}$N atom%), in SIP samples proteins may have a wide range of $^{15}$N atom%. As a result, the $^{15}$N atom% configuration of ProRata was customized for every two isotopologues of each protein. Selected ion chromatograms of identified peptides were extracted with the two $^{15}$N atom% of the two isotopologues under comparison. The selected ion chromatograms were used to estimate the peptides’ abundance ratios with principal component analysis [19], which were used to estimate the abundance ratio and its confidence interval between the two isotopologues of a protein with maximum likelihood estimation [18].
**Pathway-specific analysis of protein isotopologue abundance patterns.**

The KEGG automated annotation server (http://www.genome.jp/tools/kaas/) [28] was used to place proteins in the AMD proteomics database into KEGG orthologies using the single-directional best-hit option. Abundance ratios of protein isotopologues with less than 15 atom% $^{15}$N label incorporation to isotopologues with greater than 85 atom% $^{15}$N incorporation were used to color the KEGG Pathway maps using the KEGG API and custom scripts with green representing proteins whose light isotopologue was more abundant than the heavy isotopologue, and red for the opposite case. Proteins for which only one isotopologue was detected were assumed to have log$_2$ isotopologue abundance ratios of +/- 10.
Figures and Tables:

Figure 1: Overview of the proteomic SIP method. (A) High-resolution MS² spectrum from a shotgun proteomics measurement. The fragment ions have fully resolved isotopic distributions. (B) The MS² spectrum was matched to theoretical spectra reconstructed using predicted peptide sequences and varying ¹⁵N atom%. The optimum match to this spectrum was peptide MEGSNMTMLAPATTQAK at 49% ¹⁵N atom%. (C) ¹⁵N atom% of peptides belonging to a protein was clustered into two separate isotopologues of the protein. The abundance ratio between the two isotopologues was estimated from full-scan data using the ProRata program.

Figure 2: Validation of the proteomic SIP method. Three AMD validation samples of known ¹⁵N atom% were measured using two instruments: LTQ Orbitrap Velos and LTQ Orbitrap XL. (A) Histograms of protein ¹⁵N atom% from LTQ Orbitrap Velos analysis. (B) Histograms of protein ¹⁵N atom% from LTQ Orbitrap XL analysis. (C) Summary of validation results. False discovery rate is defined as twice of the percentage of reverse identifications out of all protein identifications. ¹⁵N atom% estimation precision is defined as the percentage of proteins within ±5% of the median ¹⁵N% of a standard sample.

Figure 3: Proteomic SIP of the AMD community regrowth. (A) Histograms of protein ¹⁵N atom%. (B) Microorganism distributions of unlabeled and labeled proteins. Each sample is represented by two bars marked by “U” or “L” for unlabeled proteins (<5% ¹⁵N atom%) and labeled proteins (>80% ¹⁵N atom%), respectively. (C) Box plots of microorganism-specific
protein abundance ratio distributions. The whiskers extend to the lowest and highest values, to a maximum distance of 1.5 times the inter-quartile range. Box plots were shown only for microorganisms with more than 5 quantified proteins. Microorganisms are color-coded in Parts B and C.

**Figure 4:** Heat maps of isotopologue abundance for the protein regrowth experiment for the four most dominant community members: three subtypes of *Leptospirillum* Bacteria (Group III, Group II UBA type, and Group II 5wayCG type), and one archaeon (G plasma of the *Thermoplasmales*). The four panels report protein identification and labeling patterns for each of the dominant organisms, although significant numbers of proteins from lower-abundance community members were detected (see text). Pale gray indicates proteins that were not identified in a specific sample. Within each panel, proteins are sorted such that those found in all samples appear at the top and those appearing in only one of the four samples appear at the bottom. The length of each panel is proportional to proteins identified. The red bar highlights a large group (234) of highly labeled proteins found only in the early regrowth sample that is enriched in proteins of no known function.

**Figure 5:** Examples of pathway-by-pathway analysis of isotopic labeling patterns as determined by proteomics analysis of dynamically $^{15}$N-labeled mixed-community biofilms. Each panel depicts protein isotopologue abundance in four different samples for proteins from a particular organism that belong to the shown pathway. The four most abundant community organisms are shown. Reductive tricarboxylic acid cycle proteins show, in general, pronounced differences in
labeling patterns even within the same organism; however, in contrast, large subunit ribosomal proteins (b) show more unified response within particular organisms.
Acknowledgements:

We thank S. Spaulding for technical assistance with cultivating AMD biofilms in laboratory bioreactors, M. Galloway for system administration of computer clusters, and N. Verberkmoes for helpful discussions. This work was funded by the US Department of Energy, Office of Biological and Environmental Research Carbon-Cycling Program (DE-SC0004665), Systems Biology Knowledgebase (DE-SC0004918) and Office of Advanced Scientific Computing Research SciDAC program. Oak Ridge National Laboratory is managed by University of Tennessee-Battelle LLC for the Department of Energy under contract DOE-AC05-00OR22725.

Author contributions:

C.P., J.F.B, C.R.F., and R.L.H. designed the research plan. D.H. and C.P. developed the data analysis algorithms and processed the data on computer clusters. C.R.F. cultivated and sampled isotopically labeled AMD biofilms. C.P. and R.L.H. designed and performed the proteomics measurements. C.P., C.R.F., and J.F.B. interpreted the results. C.P., C.R.F., and J.F.B. drafted the paper.
References:

1. M.G. Dumont and J.C. Murrell, Stable isotope probing - linking microbial identity to function. Nat Rev Microbiol, 2005. 3(6): p. 499-504.
2. H.T.S. Boschker, S.C. Nold, P. Wellsbury, D. Bos, W. de Graaf, R. Pel, R.J. Parkes, and T.E. Cappenberg, Direct linking of microbial populations to specific biogeochemical processes by 13C-labelling of biomarkers. Nature, 1998. 392: p. 801-805.
3. S. Radajewski, P. Ineson, N.R. Parekh, and J.C. Murrell, Stable-isotope probing as a tool in microbial ecology. Nature, 2000. 403(6770): p. 646-9.
4. N. Jehmlich, F. Schmidt, M. von Bergen, H.H. Richnow, and C. Vogt, Protein-based stable isotope probing (Protein-SIP) reveals active species within anoxic mixed cultures. Isme J, 2008. 2(11): p. 1122-33.
5. N. Jehmlich, F. Schmidt, M. Hartwich, M. von Bergen, H.H. Richnow, and C. Vogt, Incorporation of carbon and nitrogen atoms into proteins measured by protein-based stable isotope probing (Protein-SIP). Rapid Communications in Mass Spectrometry, 2008. 22(18): p. 2889-2897.
6. N. Jehmlich, F. Schmidt, M. Taubert, J. Seifert, M. von Bergen, H.H. Richnow, and C. Vogt, Comparison of methods for simultaneous identification of bacterial species and determination of metabolic activity by protein-based stable isotope probing (Protein-SIP) experiments. Rapid Commun Mass Spectrom, 2009. 23(12): p. 1871-8.
7. N. Jehmlich, I. Fetzer, J. Seifert, J. Mattow, C. Vogt, H. Harms, B. Thiede, H.H. Richnow, M. von Bergen, and F. Schmidt, Decimal place slope, a fast and precise method for quantifying 13C incorporation levels for detecting the metabolic activity of microbial species. Mol Cell Proteomics, 2010. 9(6): p. 1221-7.
8. I. Fetzer, N. Jehmlich, C. Vogt, H.H. Richnow, J. Seifert, H. Harms, M. von Bergen, and F. Schmidt, Calculation of partial isotope incorporation into peptides measured by mass spectrometry. BMC Res Notes, 2010. 3: p. 178.
9. F. Bastida, M. Rosell, A.G. Franchini, J. Seifert, S. Finsterbusch, N. Jehmlich, S. Jechalke, M. von Bergen, and H.H. Richnow, Elucidating MTBE degradation in a mixed consortium using a multidisciplinary approach. FEMS Microbiol Ecol, 2010. 73(2): p. 370-84.
10. R.J. Ram, N.C. Verberkmoes, M.P. Thelen, G.W. Tyson, B.J. Baker, R.C. Blake, 2nd, M. Shah, R.L. Hettich, and J.F. Banfield, Community proteomics of a natural microbial biofilm. Science, 2005. 308(5730): p. 1915-20.
11. C.P. Belnap, C. Pan, N.C. Verberkmoes, M.E. Power, N.F. Samatova, R.L. Carver, R.L. Hettich, and J.F. Banfield, Cultivation and quantitative proteomic analyses of acidophilic microbial communities. Isme J, 2009.
12. I. Lo, V.J. Denef, N.C. Verberkmoes, M.B. Shah, D. Goltsman, G. DiBartolo, G.W. Tyson, E.E. Allen, R.J. Ram, J.C. Detter, P. Richardson, M.P. Thelen, R.L. Hettich, and J.F. Banfield, Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature, 2007. 446(7135): p. 537-41.
13. V.J. Denef, R.S. Mueller, and J.F. Banfield, AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. Isme J, 2010.
14. G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield, Community structure and
metabolism through reconstruction of microbial genomes from the environment. Nature, 2004. 428(6978): p. 37-43.

15. M.P. Washburn, D. Wolters, and J.R. Yates, 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol, 2001. 19(3): p. 242-7.

16. A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, and S. Horning, Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal Chem, 2006. 78(7): p. 2113-20.

17. J.V. Olsen, J.C. Schwartz, J. Grieb-Raming, M.L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E.R. Wouters, M. Senko, A. Makarov, M. Mann, and S. Horning, A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics, 2009. 8(12): p. 2759-69.

18. C. Pan, G. Kora, W.H. McDonald, D.L. Tabb, N.C. VerBerkmoes, G.B. Hurst, D.A. Pelletier, N.F. Samatova, and R.L. Hettich, ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. Anal Chem, 2006. 78(20): p. 7121-31.

19. C. Pan, G. Kora, D.L. Tabb, D.A. Pelletier, W.H. McDonald, G.B. Hurst, R.L. Hettich, and N.F. Samatova, Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. Anal Chem, 2006. 78(20): p. 7110-20.

20. P. Wilmes, J.P. Remis, M. Hwang, M. Auer, M.P. Thelen, and J.F. Banfield, Natural acidophilic biofilm communities reflect distinct organismal and functional organization. Isme J, 2009. 3(2): p. 266-70.

21. D.S. Goltsman, V.J. Denef, S.W. Singer, N.C. VerBerkmoes, M. Lefsrud, R.S. Mueller, G.J. Dick, C.L. Sun, K.E. Wheeler, A. Zemla, B.J. Baker, L. Hauser, M. Land, M.B. Shah, M.P. Thelen, R.L. Hettich, and J.F. Banfield, Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms. Appl Environ Microbiol, 2009. 75(13): p. 4599-615.

22. K.J. Edwards, P.L. Bond, T.M. Gihring, and J.F. Banfield, An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. Science, 2000. 287(5459): p. 1796-9.

23. B.J. Baker, G.W. Tyson, R.I. Webb, J. Flanagan, P. Hugenholtz, E.E. Allen, and J.F. Banfield, Lineages of acidophilic archaea revealed by community genomic analysis. Science, 2006. 314(5807): p. 1933-5.

24. B.J. Baker, L.R. Comolli, G.J. Dick, L.J. Hauser, D. Hyatt, B.D. Dill, M.L. Land, N.C. Verberkmoes, R.L. Hettich, and J.F. Banfield, Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A, 2010. 107(19): p. 8806-11.

25. N.C. VerBerkmoes, V.J. Denef, R.L. Hettich, and J.F. Banfield, Systems biology: Functional analysis of natural microbial consortia using community proteomics. Nat Rev Microbiol, 2009. 7(3): p. 196-205.

26. C. Pan, B.H. Park, W.H. McDonald, P.A. Carey, J.F. Banfield, N.C. Verberkmoes, R.L. Hettich, and N.F. Samatova, A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. BMC Bioinformatics. 11(1): p. 118.
27. J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, and S.P. Gygi, *Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome*. J Proteome Res, 2003. **2**(1): p. 43-50.

28. Y. Moriya, M. Itoh, S. Okuda, A.C. Yoshizawa, and M. Kanehisa, *KAAS: an automatic genome annotation and pathway reconstruction server*. Nucleic Acids Res, 2007. **35**(Web Server issue): p. W182-5.
**Figure 1**

(A) Orbitrap MS$^2$ spectrum of a 50% $^{15}$N-labeled peptide

![Zoom-in](image)

Isotopically resolved fragment ions

(B) Peptides predicted by *in silico* digestion

| Stable isotope Incorporation level | Peptide 1 | Peptide 2 | Peptide 3 | ..... | Peptide N |
|-----------------------------------|-----------|-----------|-----------|-------|-----------|
| 0%                                | 2-Dimensional MS/MS Identification | Search Space |
| 1%                                |           |           |           |       |           |
| 2%                                |           |           |           |       |           |
| .....                              |           |           |           |       |           |
| 100%                              |           |           |           |       |           |

(C) Isotopologue 1 Isotopologue 2

Proteins

Peptides

Clustering

| 15N Atom% | 0% | 100% |
|-----------|----|------|
| Isotopologue 1 | Peptides | Proteins |
| Isotopologue 2 | Peptides | Peptides |
Figure 2

(A) LTQ Orbitrap Velos
(B) LTQ Orbitrap XL

| MS Instrument Model | LTQ Obitrap Velos | LTQ Orbitrap XL |
|---------------------|------------------|-----------------|
| Expected $^{15}$N atom% of samples | 0.4% 50% 98% | 0.4% 50% 98% |
| Non-redundant AMD proteins | 2326 1408 1972 | 1281 815 1237 |
| False discovery rate | 1.02% 1.96% 0.81% | 0.47% 0.97% 0% |
| Median estimated $^{15}$N atom% | 0% 48% 98% | 1% 48% 98% |
| $^{15}$N atom% estimation precision | 99.9% 94.6% 99.7% | 99.9% 96.6% 99.8% |
Protein Counts

Sample A1

Sample A2

Sample B1

Sample B2

Established Biofilm

Regrowing Biofilm

Time: 105 hrs

Time: 274 hrs

(A)

(B)

(C)

Microorganisms:
- Leptospirillum III
- Others
- G-plasma
- Ferroplasma II
- Ferroplasma I
- Lepto.II NU
- Lepto.II UBA
- Lepto.II 5wayCG
Figure 4

Leptospirillum group III

Leptospirillum group II SwayCG

Leptospirillum group II UBA

G Plasma

\( ^{15}\text{N-rich} \quad \longleftrightarrow \quad ^{14}\text{N-rich} \)

Isotopologue Abundance Ratio, log\(_2\) scale
Figure 5B