Published in final edited form as:

Trends Ecol Evol. 2012 August ; 27(8): 436–442. doi:10.1016/j.tree.2012.03.001.

Environmental Proteomics, Biodiversity Statistics, and Food-Web Structure

Nicholas J. Gotelli^{1,*}, Aaron M. Ellison², and Bryan A. Ballif^{1,3}

¹Department of Biology, University of Vermont, 109 Carrigan Drive, Burlington, VT 05405, USA

²Harvard Forest, Harvard University, 324 North Main Street, Petersham, MA 01366, USA

³Vermont Genetics Network Proteomics Facility, University of Vermont, 109 Carrigan Drive , Burlington, VT 05405, USA

Abstract

Pioneering studies in environmental proteomics have revealed links between protein diversity and ecological function in simple ecological communities such as microbial biofilms. In the near future, high throughput proteomic methods will be applied to more complex ecological systems in which microbes and macrobes interact. Data structures in biodiversity and protein surveys have many similarities, so the statistical methods that ecologists use for analyzing biodiversity data should be adapted for use with quantitative surveys of protein diversity. However, increasing quantities of protein and bioinformatics data will not, by themselves, reveal the functional significance of proteins. Instead, ecologists should be measuring changes in the abundance of protein cohorts in response to replicated field manipulations, including nutrient enrichment and removal of top predators.

SURVEYING THE PROTEIN DIVERSITY OF COMMUNITIES AND ECOSYSTEMS

High throughput methods for studying and characterizing large numbers of proteins of uncultured biological samples (metaproteomics) now allow for cataloguing the proteins of component species of assemblages (community proteomics) or of parts of ecosystems, without precise knowledge of the organism that produced the protein (environmental proteomics) [1,2]. Pioneering studies have surveyed relatively simple, microbial-dominated assemblages (reviewed in [3,4]), sampled from habitats such as acid-mine drainages [5], microfilm surfaces of leaves [6], seawater [7] and soil [8] (Box 1).

These methods and results should be of great interest to all community and ecosystem ecologists, because patterns of differential protein production and expression reflect physiological responses to changing or stressful environmental conditions [9], including climatic change [10] and the presence of predators [11]. In addition, shifts in protein abundance and composition — best characterized through analyses of microbial processes [12] — can indicate changes in magnitude or rates of material and energetic fluxes within and between ecosystems [13].

© 2012 Elsevier Ltd. All rights reserved.

*Correspondence Author (ngotelli@uvm.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Although pioneering studies of microbes suggest directions for future work in environmental proteomics, it is unclear how easily results from microbial assemblages can be scaled to more complex ecological systems. The structural complexity of microbial systems appears to be simplified relative to that of most terrestrial and aquatic ecosystems. Biofilm and microbial systems generally lack the photosynthetic primary producers, detritivores, and higher trophic levels that characterize typical “green” and “brown” food webs [14] of interacting multicellular (“macrobial”) eukaryotes and prokaryotes [15]. Technical challenges—including the identification of proteins from species without sequenced genomes, variation in the physical and biological conformation of proteins, *in situ* activity of isolated proteins and the difficulty of reliably extracting proteins from complex media such as seawater and soil [16] — also hinder the application of proteomic (and genomic) approaches to macrobial assemblages and food webs [3]. Nevertheless, as these challenges are overcome with new technologies and bioinformatics tools, proteomic surveys of a range of ecological communities and ecosystems will become an important complement to parallel metagenomic, metatranscriptomic and metametabolomic surveys [2]. All of these methods generate complementary data, although from our perspective, proteins are the most desirable unit of study because they are most closely related to the functioning of ecosystems and because they are a more direct measure of the “molecular phenotype” [17].

Perhaps the most important practical result of ongoing technological advances is that ecologists will be able to survey proteomes efficiently and repeatedly and describe their variability in natural ecosystems, rather than ignoring variation and classifying species, communities and their proteomes using fixed typologies. For example, Figure 1 shows three proteomic “fingerprints” of samples from three replicate food webs inhabiting water-filled leaves of the pitcher plant *Sarracenia purpurea*. Each pitcher supports an independent, intact food web consisting of a resource base of captured insect prey (mostly ants and flies), a sub-web of bacteria, protozoa, algae, and fungi and a suite of aquatic invertebrate larvae (flies, midges and mosquitoes) that interact as filter feeders, detritivores, omnivores, and top predators [18]. The obvious variability in protein profiles among replicate food webs in the microbial component of each web (Figure 1) likely reflects differences in species composition and abundance of invertebrates and microbes within each leaf, along with differences in the quantity and composition of captured insect prey. At the same time, consistent protein bands suggest repeatable patterns and protein structure among replicate food webs. In an associated community proteomics survey [19], species-specific proteins and peptide signatures of the three most abundant macroinvertebrates were revealed, a first step towards proteomic profiling of an entire microbe-macrobe ecosystem.

UNDERSTANDING THE FUNCTIONAL SIGNIFICANCE OF PROTEIN DIVERSITY

In just two decades, cost-effective methods have been developed to increase throughput and comprehensive identification of proteins [20,21]; accurately quantify protein relative abundance [22,23]; enrich rare protein types and protein modifications [24, 25]; and expand the capacity and precision of bioinformatic tools that enhance the functional interpretation of proteomic datasets [26,27]. Although some of these methods are still in their infancy, rapid technological and statistical advances daily are increasing the quantity and quality of information about proteins that can be sampled from functioning ecosystems. However, by itself, this flood of bioinformatics data is unlikely to reveal which particular proteins are important causes and consequences of community structure and ecosystem function.

An analogous disconnection between data density and functional inference plagued community ecology in the initial stages of documenting the diversity of species, their assembly, and their interactions within food webs. Pioneering naturalists and taxonomists —

Rumphius, Linneaus, Darwin, Wallace, and von Humboldt, among many others —collected, catalogued and named species from around the world. Beginning in the late 1800s, ecologists assembled detailed records of food web structure —extensive lists of “who eats whom” — in plant and animal assemblages. Now, species diversity and food web structure both are encapsulated, respectively, in summary statistics and indices [28] and as network diagrams with nodes representing species and links representing trophic interactions including predation and parasitism [29].

But pattern identification and summary statistics are not the same as identifying causal processes, which requires experimental manipulations and statistical modeling of the results. In studies of biodiversity [30,31] and ecological food webs [32,33], there are long dialectical traditions of thesis, antithesis, and synthesis. The same dialectic could be applied profitably to proteomics research in ecology: characterizing and summarizing the diversity of proteins and their potential interactions, analyzing proteomic assemblages of experimentally manipulated food webs in the field and synthesizing the resulting data with statistical modeling.

CHARACTERIZING AND SUMMARIZING PROTEOMIC DIVERSITY

Just as the first step to working with an ecosystem is identifying its constituent organisms, a first step towards understanding the functional significance of protein diversity of an ecosystem will be identifying and enumerating their distinct proteins and estimating their relative abundances. However, just as it is impossible to detect all of the species in a microbial or macrobial assemblage [34,35], so too it is unlikely that we will be able to detail an entire ecological proteome in the foreseeable future. Estimates of proteomic diversity in microbial ecosystems range from 10^4 to 10^9 expressed proteins [16]. There are approximately 6,000–60,000 protein-encoding genes in a single prokaryote or eukaryote species' genome [36], and even a *single* tissue sampled from a *single* multicellular species contains hundreds or even thousands of distinct proteins [37]. In ecological assemblages that contain both microbes and macrobes, proteomic diversity likely will exceed 10^{10} expressed proteins because— unlike the — genome the proteome of a multicellular organism is not constant during its lifetime, but changes ontogenetically [38] and in response to changing biotic and abiotic conditions [9].

For proteins, as for species, ecologists must remain satisfied with sampling only a small fraction of the potential diversity. The resulting data from surveys of both ecosystems and proteomes are lists of the component species (some of which might be taxonomically ambiguous) or proteins (some of which might be incompletely characterized because a single tryptic peptide might map to several, typically related, proteins) and an estimate of relative abundances (usually from counts of individual organisms [31] or counts of peptides from different proteins [39]). Statistical distributions of both species and proteins are also similar: there are usually a small number of relatively common species (or proteins), a large number of relatively rare species (or proteins) and an unknown number of species (or proteins) that are present, but are not detected in the sample. The observed distribution of relative abundances of species is roughly log-normal in shape [40], although the precise statistical characterization continues to be controversial [41]. However, the observed distribution of the relative abundance of proteins has not yet been analyzed with the same statistical tools that ecologists have used to analyze the relative abundance of species.

Over the past several decades, ecologists have developed a unified toolbox of non-parametric statistical methods for analyzing such data. Rarefaction methods interpolate diversity estimates of small random subsamples of data, allowing investigators to compare multiple, standardized data sets on the basis of common sampling effort [42]. Asymptotic

estimators can extrapolate diversity data to estimate the minimum number of undetected species (or proteins) in a sample, based on the frequencies of the rare species that are present [43]. The amount of sampling effort that would be needed to reveal these undetected elements also can be calculated [44]. Finally, sample variances and confidence intervals characterize the uncertainty associated with such interpolations and extrapolations [45].

These statistical tools have not yet been adapted for protein studies; Koziol *et al.* [46] is an exemplary study, and the only published example we know of in which biodiversity sampling models were applied to proteomic data (Figure 2). Although it might seem premature to discuss statistical methods for data that have not yet been collected at the ecosystem and community level, we think it is critical to look forward and anticipate methods of analysis for proteomic data that will emerge in future studies. Such analyses might reveal, for example, whether the relationship between protein diversity and ecosystem function mirrors that of species diversity and ecosystem function.

Analysis of proteomic data differs from analysis of biodiversity samples in one crucial way, however. To characterize a proteome, the proteins are proteolytically digested into smaller peptides prior to analysis; this is analogous to taking the species sampled, shredding them into small pieces, and then re-assembling, identifying, and counting the reconstituted organisms! Considerable efforts in bioinformatics are devoted towards statistically reconstructing the original protein(s) from these peptide fragments. This added twist makes it difficult to apply traditional analyses of random samples of “individuals” to the estimation of relative abundances of different protein types. In particular, standard rarefaction of peptide frequency data will be biased: assuming a similar distribution in the size and ionization efficiency of peptides from each protein, the frequency of large proteins will be overestimated and the frequency of small proteins will be underestimated. However, using the number of amino acids in each identified protein, a simple modification to rarefaction is possible that will help to correct such biases (Figure 3).

Proteomics data, like biodiversity data, increasingly are available over the Internet from institutional data repositories such as the Proteomics Identifications database (PRIDE <http://www.ebi.ac.uk/pride>). Unlike ecological data, however, proteomic data come in only a few, highly standardized forms – mass-spectroscopy spectra and protein identification supported by peptide identifications – that are routinely (and mandatorily) submitted to data archives as part of the manuscript submission and publication process. Proteomic data repositories require detailed metadata about the origins, processing, and analysis of samples, and they have controlled vocabularies (ontologies) for annotating data and metadata that can be searched with predefined computational algorithms and structured work-flows [47]. Thus, these data repositories provide an additional, as yet untapped, resource for exploring patterns of proteomic diversity.

In summary, there are remarkable similarities between the biodiversity data and environmental and community proteomics data, in the constraints and challenges of sampling, in the form of the data that result from such sampling, and in the availability of archived data and work-flows for analyzing them. With adjustments to account for bias in protein size, the statistical framework developed by ecologists to characterize and summarize biological diversity will be applicable to protein diversity data.

EXPERIMENTAL ENVIRONMENTAL PROTEOMICS

Summary metrics and statistical analyses are only the first step forward with proteomic data. We are more interested in how proteins interact in an assemblage of organisms or in a complete ecosystem, and expect that an understanding of these interactions will provide new insights into the workings of ecological processes. As noted earlier, the sheer number of

proteins in even a single tissue suggests that trying to identify each and every function for each and every protein in an ecological system will be a long time coming, and might not be informative for ecologists working with typical food webs of macrobes and microbes. Mechanistic studies of food webs have grappled with the analogous question – do we need to know the identity and role of every species in a food web, or can we reduce the inherent complexity of food webs by grouping species into trophic categories (consumers, detritivores, predators), functional feeding groups (shredders, scrapers, filter-feeders), or taxonomic/ecological guilds (tube-building polychaetes, seed-eating finches) that still provide robust insights into mechanisms driving ecological dynamics? When these simplified groupings have been organized into abstracted webs that have few nodes and links [48], they have proven to be more amenable to mechanistic modeling [49]. The predictions of these simplified models have been tested directly by adding or removing species at different trophic levels [50] manipulating nutrients or basal resources [51], or quantifying the responses of the rest of the assemblage to such perturbations [52].

A similar approach might profitably be used to understand the functional significance of proteins in ecological systems. This would be achieved most easily by grouping proteins (or their homologs) by class using gene ontology databases in which proteins are classified into functional categories based on similar biological processes, biochemical activities, and subcellular localization [53, 54]. Indeed, one of the key reasons that proteomics might yield new ecological insights is that many species —particularly of microbes —might be functionally equivalent because they produce similar proteins with similar functions in similar environments [55,56]. This kind of ecological redundancy is a key concept in the study of ecosystem function [57] that already allows ecologists to simplify the vast diversity that characterizes natural assemblages [58].

Thus, rather than trying to characterize all of the uncommon proteins in an assemblage, ecologists should first concentrate on the small number of identified proteins that change in abundance either from common to rare or rare to common in response to experimental manipulations. By adding or removing species and species groups, ecologists will discover which kinds of proteins are involved in the response. This molecular phenotyping [17] of entire assemblages should provide insight into how these proteins function, and repeated sampling might reveal temporal dynamics as well. Molecular phenotyping and recognizing proteins' functions do not require that functionally important proteins be common in natural assemblages. When proteins that are absent or rare in control assemblages become abundant in response to experimental perturbations, there is good reason to suspect they have an important function in the ecosystem.

Moreover, as recombinant DNA technology and protein expression systems become widely available to ecologists, entirely new avenues of ecological experiments will become possible, at least for secreted proteins. Rather than manipulating species or trophic groups, ecologists will be able to add synthesized proteins directly to experimental mesocosms. There is a long tradition of experimentally (or unintentionally) enriching food webs with nitrogen, phosphorous, and other critical nutrients that limit the growth of phytoplankton, algae, and terrestrial plants [59,60]. This idea can be extended to enriching ecosystems with potentially important proteins and metabolites, and then measuring food-web responses. Measuring the ecological proteome in response to species removals and additions, and measuring the biodiversity response to the addition of synthetic proteins should go lead ecologists to understand the functions of proteins in ecosystem organization.

Of course, these kinds of experiments will generate new statistical challenges. For example, if 500 proteins are surveyed in an ecological field experiment with a control and a predator removal treatment, a traditional frequentist analysis would potentially generate 500 *t*-tests

and associated *P*-values. Even if there are no effects of the top predator on the ecological proteome, at least some of these comparisons will be statistically significant by chance alone [61]. However, promising new analyses based on empirical Bayesian [62] and other approaches [63] have been effective in screening large data sets such as those generated by proteomic surveys [64], microarrays [65] and species occurrence matrices [66].

SUMMARY

Pioneering studies of environmental proteomics have recently revealed the functional significance of proteins in simple biofilm communities. However, it is not clear how well the results can be generalized to more typical communities of interacting microbes and macrobes in structurally complex environments. Technological advances and improvements in bioinformatics and data archiving will allow for proteomic characterization of a variety of terrestrial and aquatic habitats. However, protein diversity is vast, and there will always be undetected rare proteins in biological material. Rather than trying to exhaustively sample this rare tail of protein diversity, ecologists should concentrate on measuring more common proteins in replicated samples, and quantifying the variability inherent in natural food webs.

If the relative abundances of different protein types from a biological sample can be estimated, the resulting data are remarkably similar in form to counts of individuals and species that already are familiar to ecologists who analyze patterns of biodiversity. For such analyses, a statistical tool box of rarefaction methods for interpolation, and asymptotic estimators for extrapolation can be applied to effectively summarize protein diversity data and will allow ecologists to standardize samples for meaningful comparisons.

However, as the history of food web studies has illustrated, statistical methods alone usually cannot reveal important functional roles. Therefore, ecologists should begin using proteomics in combination with traditional experimental field manipulations, such as nutrient enrichment, species additions and removals, and the modification of habitat complexity. In such experiments, proteins that change status from “common” to “rare” or “rare” to “common” are good candidates for functional importance to food web structure, although in some cases they might simply mirror changes in the relative abundance of constituent species. More innovative experiments would involve the addition to food webs of synthetic proteins first identified by proteomic analysis.

Although there are still serious methodological challenges to extracting proteins from complex media such as soil and seawater, proteomics will eventually become an important tool for community and ecosystem ecologists. The analysis of environmental proteomic data—in combination with statistical biodiversity methods and experimental field manipulations—has the potential to greatly increase our understanding of the role of proteins in the organization and function of food webs and ecosystems.

Acknowledgments

We thank Rachel Brooks for generating the data in Figure 1. N.J.G. was supported by the U.S. National Science Foundation (DEB-136703 and DEB-0541936) and the U.S. Department of Energy (DE-FG02-08ER64510), and a University of Vermont Research Award to N.J.G. and B.A.B. A.M.E. was supported by grants from the U.S. National Science Foundation (DEB-136703, DEB-0541680, DEB 06-20443, and DBI-1003938), and the U.S. Department of Energy (DE-FG02-08ER64510). B.A.B. was also supported by the U.S. National Science Foundation (DEB-136703) and the Vermont Genetics Network through U.S. National Institutes of Health grant P20 RR16462 from the INBRE Program of the NCRR.

GLOSSARY

community proteomics	Proteomic surveys of the taxa that comprise a local assemblage or community
environmental proteomics	Proteomic surveys of an assemblage or community without specification of the particular taxon that a protein originated from
gene ontology database	A relational database categorizing genes and proteins by function, expression level, biochemical activity, metabolic process, subcellular localization, or similar attributes for bioinformatic analyses
metagenomics	The large-scale characterization of genetic material recovered from a biological or environmental source
metametabolomics	The large-scale characterization of small-molecule metabolites from a biological or environmental source; <i>metaproteomics</i> The large-scale characterization of proteins recovered from a biological or environmental source.
metatranscriptomics	The large-scale characterization of transcribed RNA molecules from a biological or environmental source; <i>molecular phenotyping</i> The large-scale composite characterization of biological samples by surveying differences in transcript or protein levels.
tryptic peptide	The proteolytic product following the digestion of a protein with trypsin. Trypsin is a reliable and economical enzyme that cuts after basic residues in proteins, yielding peptides that are highly compatible with mass spectrometry applications

LITERATURE CITED

1. Maron PA, et al. Metaproteomics: A new approach for studying functional microbial ecology. *Microbial Ecology*. 2007; 53:486–493. [PubMed: 17431707]
2. Schneider T, Riedel K. Environmental proteomics: Analysis of structure and function of microbial communities. *Proteomics*. 2010; 10:785–798. [PubMed: 19953545]
3. Vandenkoornhuysen P, et al. Integration of molecular functions at the ecosystemic level: breakthroughs and future goals of environmental genomics and post-genomics. *Ecology Letters*. 2010; 13:776–791. [PubMed: 20426792]
4. VerBerkmoes NC, et al. Functional analysis of natural microbial consortia using community proteomics. *Nature Reviews Microbiology*. 2009; 7:196–205.
5. Ram RJ, et al. Community proteomics of a natural microbial biofilm. *Science*. 2005; 308:1915–1920. [PubMed: 15879173]
6. Delmotte N, et al. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:16428–16433. [PubMed: 19805315]
7. Kan J, et al. Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems*. 2005; 1:7. [PubMed: 16176596]
8. Schulze WZ, et al. A proteomics fingerprint of dissolved organic carbon and of soil particles. *Oecologia*. 2005; 142:335–343. [PubMed: 15449171]
9. Nesatyy VJ, Suter MJF. Proteomics for the analysis of environmental stress responses in organisms. *Environmental Science & Technology*. 2007; 41:6891–6900. [PubMed: 17993125]

10. Tomanek L, Zuzow MJ. The proteomic response of the mussel congeners *Mytilus galloprovincialis* and *M. trossulus* to acute heat stress: implications for thermal tolerance limits and metabolic costs of thermal stress. *Journal of Experimental Biology*. 2010; 213:3559–3574. [PubMed: 20889836]
11. Vaitkevicius K, et al. A *Vibrio cholerae* protease needed for killing of *Caenorhabditis elegans* has a role in protection from natural predator grazing. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103:9280–9285. [PubMed: 16754867]
12. Banfield JF, et al. Proteogenomic approaches for the molecular characterization of natural microbial communities. *Omics-a Journal of Integrative Biology*. 2005; 9:301–333. [PubMed: 16402891]
13. Schultze W. Environmental proteomics – what proteins from soil and surface water can tell us: a perspective. *Biogeosciences Discussions*. 2004; 1:195–218.
14. Grossert HP. Ecological consequences of bacterioplankton lifestyles: changes in concepts are needed. *Environmental Microbiology Reports*. 2010; 2:706–714.
15. Moore JC, et al. Detritus, trophic dynamics and biodiversity. *Ecology Letters*. 2004; 7:584–600.
16. Wilmes P, Bond PL. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends in Microbiology*. 2006; 14:92–97. [PubMed: 16406790]
17. Biron DG, et al. Population proteomics: an emerging discipline to study metapopulation ecology. *Proteomics*. 2006; 6:1712–1715. [PubMed: 16429464]
18. Gotelli NJ, Ellison AM. Food-web models predict species abundances in response to habitat change. *Plos Biology*. 2006; 4:1869–1873.
19. Gotelli NJ, et al. Proteomic characterization of the major arthropod associates of the carnivorous pitcher plant *Sarracenia purpurea*. *Proteomics*. 2011; 11:2354–2358. [PubMed: 21538880]
20. Beck M, et al. Comprehensive proteomics. *Current Opinion in Biotechnology*. 2011; 22:3–8. [PubMed: 20888217]
21. Thakur SS, et al. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & Cellular Proteomics*. 2011:10.
22. Wilm M. Quantitative proteomics in biological research. *Proteomics*. 2009; 9:4590–4605. [PubMed: 19743428]
23. Walther TC, Mann M. Mass spectrometry-based proteomics in cell biology. *Journal of Cell Biology*. 2010; 190:491–500. [PubMed: 20733050]
24. Groen AJ, Lilly KS. Proteomics of total membranes and subcellular membranes. *Expert Reviews of Proteomics*. 2010; 7:867–878.
25. Goswami, T.; Ballif, BA. Methods for the isolation of phosphoproteins and phosphopeptides for mass spectrometry analysis: Toward increased functional phosphoproteomics. In: Ivanov, A.; Lazarev, A., editors. *Sample Preparation in Biological Mass Spectrometry*. Springer; 2011. p. 627–655.
26. Chen C, et al. Protein bioinformatics databases and resources. *Methods in Molecular Biology*. 2011; 694:3–24. [PubMed: 21082424]
27. Jacob RJ. Bioinformatics for LC-MS/MS-based proteomics. *Methods in Molecular Biology*. 2010; 658:61–91. [PubMed: 20839098]
28. Jost L. Partitioning diversity into independent alpha and beta components. *Ecology*. 2007; 88:2427–2439. [PubMed: 18027744]
29. Digel C, et al. Body sizes, cumulative and allometric degree distributions across natural food webs. *Oikos*. 2011; 120:503–509.
30. Rosenzweig, M. *Species Diversity in Space and Time*. Cambridge University Press; 1995.
31. Magurran, AE.; McGill, BJ., editors. *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford; 2011.
32. Cohen, JE. *Food webs and Niche Space*. Princeton University Press; 1978.
33. Pascual, M.; Dunne, JA., editors. *Ecological Networks: Linking Structure to Dynamics in Food Webs*. Oxford University Press; 2005.
34. Hughes JB, et al. Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*. 2001; 67:4399–4406. [PubMed: 11571135]

35. Longino JT, et al. The ant fauna of a tropical rain forest: Estimating species richness three different ways. *Ecology*. 2002; 83:689–702.
36. Van Straalen, NI.; Roelofs, D. Introduction to Ecological Genetics. Oxford University Press; 2006.
37. Shen YF, et al. Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Analytical Chemistry*. 2004; 76:1134–1144. [PubMed: 14961748]
38. Guercio RAP, et al. Ontogenetic variations in the venom proteome of the Amazonian snake *Bothrops atrox*. *Proteome Science*. 2006;4. [PubMed: 16571109]
39. Lundgren DH, et al. Role of spectral counting in quantitative proteomics. *Expert Review of Proteomics*. 2010; 7:39–53. [PubMed: 20121475]
40. McGill, BJ. Species abundance distributions. In: Magurran, AE.; McGill, BJ., editors. *Biological Diversity: Frontiers In Measurement & Assessment*. Oxford; 2011. p. 105-122.
41. Connolly SR, et al. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology*. 2009; 90:3138–3149. [PubMed: 19967869]
42. Gotelli, NJ.; Colwell, RK. Estimating species richness. In: Magurran, AE.; McGill, BJ., editors. *Biological Diversity: Frontiers In Measurement & Assessment*. Oxford University Press; 2011. p. 39-54.
43. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*. 1994; 345:101–118.
44. Chao A, et al. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*. 2009; 90:1125–1133. [PubMed: 19449706]
45. Colwell RK, et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblages. *Journal of Plant Ecology*. 2012; 5:3–21.
46. Koziol JA, et al. Application of capture-recapture models to estimation of protein count in MudPIT experiments. *Analytical Chemistry*. 2006; 78:3203–3207. [PubMed: 16643015]
47. Michener WK, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*. 2012; 27:85–93. [PubMed: 22240191]
48. Holyoak M, Sachdev S. Omnivory and the stability of simple food webs. *Oecologia*. 1998; 117:413–419.
49. Allesina S, et al. A general model for food web structure. *Science*. 2008; 320:658–661. [PubMed: 18451301]
50. Wootton JT. Indirect effects, prey susceptibility, and habitat selection - impacts of birds on limpets and algae. *Ecology*. 1992; 73:981–991.
51. Becks L, et al. Experimental demonstration of chaos in a microbial food web. *Nature*. 2005; 435:1226–1229. [PubMed: 15988524]
52. Eveleigh ES, et al. Fluctuations in density of an outbreak species drive diversity cascades in food webs. *Proceedings of the National Academy of Sciences of the United States of America*. 2007; 104:16976–16981. [PubMed: 17940003]
53. Lacerda CMR, et al. Metaproteomic analysis of a bacterial community response to cadmium exposure. *Journal of Proteome Research*. 2007; 6:1145–1152. [PubMed: 17284062]
54. Matsuoka S, et al. ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*. 2007; 316:1160–1166. [PubMed: 17525332]
55. Nagy R, et al. The characterization of novel mycorrhiza-specific phosphate transporters from *Lycopersicon esculentum* and *Solanum tuberosum* uncovers functional redundancy in symbiotic phosphate transport in solanaceous species. *Plant Journal*. 2005; 42:236–250. [PubMed: 15807785]
56. Schloss PD, Handelsman J. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *Bmc Bioinformatics*. 2008;9. [PubMed: 18182098]
57. Rosenfield JS. Functional redundancy in ecology and conservation. *Oikos*. 2002; 98:156–162.
58. Mouchet MA, et al. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*. 2010; 24:867–876.

59. Hecky RE, Kilham P. Nutrient limitation of phytoplankton in fresh-water and marine environments - a review of recent evidence on the effects of enrichment. *Limnology and Oceanography*. 1988; 33:796–822.
60. Smith VH, et al. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environmental Pollution*. 1999; 100:179–196. [PubMed: 15093117]
61. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*. 1995; 57:289–300.
62. Efron B. The future of indirect evidence. *Statistical Science*. 2010; 25:145–157. [PubMed: 21243111]
63. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 106:9440–9445. [PubMed: 12883005]
64. Huttlin EL, et al. Prediction of error associated with false positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *Journal of Proteome Research*. 2007; 6:392–398. [PubMed: 17203984]
65. Dudoit S, et al. Multiple hypothesis testing in microarray experiments. *Statistical Science*. 2003; 18:71–103.
66. Gotelli NJ, Ulrich W. The empirical Bayes approach as a tool to identify non-random species associations. *Oecologia*. 2010; 162:463–477. [PubMed: 19826839]
67. Durr E, et al. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nature Biotechnology*. 2004; 22:985–992.
68. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*. 1987; 43:783–791. [PubMed: 3427163]
69. Baker BJ, Banfield JF. Microbial communities associated with acid mine drainage. *FEMS Microbial Review*. 2003; 44:139–152.

Box 1**Omic characterization of community structure and function**

Acid-rock and acid-mine drainages (AMDs) can be found wherever water moves through pyrite (iron sulfide: FeS_2); examples include mineral-rich deposits in caves and piles of tailings from locations where sulfide minerals (including uranium and coal) are mined. AMDs have very low pH (often reaching 0.5–1.0), but can host highly productive ecosystems consisting of diverse bacteria, Archaea, and eukaryotes. Many of these taxa are chemoautotrophs that have evolved adaptations to heavy-metal environments [69]. The chemical isolation of these communities from external sources of carbon and nitrogen, and their physical segregation from higher pH assemblages in specific drainages makes them an ideal self-contained, microecosystem for assessing proteomic diversity and function in a microbial system.

Banfield *et al.* [12] sampled the AMD biofilm community at six sites along an environmental gradient (defined by distance from ore deposits and seepage flow) at the Richmond Iron Mountain Mine Complex in northern California. Diversity of microbes from water samples and biofilm scraped from rock surfaces was determined using a combination of in situ hybridization with fluorescent probes screened with epifluorescence and confocal microscopy, followed by direct DNA sequencing of entire genomes. The sequence data were then used to identify proteins from particular dominant organisms. Unknown sequences were screened against proteomic databases to identify common proteins. Functions of novel proteins were characterized further with biological observation, mass-spectroscopy, biochemical assays (including microtiter plates that can assay ~100,000 samples at a time for multiple biochemical functions) and protein-structural modeling.

Figure 1.

SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) separation of proteins from the microbial fraction of material within three individual *Sarracenia purpurea* ecosystems. The pitcher-plant liquid was extracted from each plant and filtered using a Bio-Rad polyprep column with a porous 30 μm polyethylene frit bed and gentle suction. 2 ml of the filtered liquid was subjected to centrifugation at $\sim 13,000 \times g$. The pellet was re-suspended in 100 μl denaturing protein sample buffer, heated to 95 $^{\circ}\text{C}$ for five minutes, and 20 μl was subjected to SDS-PAGE. Gels were stained with Coomassie blue. Approximate molecular weights are indicated in kilodaltons (kDa). Some protein bands appear to be

common to all samples (e.g., ~55 kDa), whereas others (e.g., ~82 kDa and ~140 kDa in Sample 1) appear to be unique to a particular sample.

Figure 2.

Sample-based rarefaction analysis of proteomic data. The x -axis is the number of MudPIT (multidimensional protein identification technology) samples of rat endothelial tissue (original data from [67]). The y -axis is the number of unique proteins discovered. The set of 10 MudPIT samples yielded a total of 426 unique proteins (left solid circle). The solid red line is the rarefaction curve interpolated from the original data to smaller numbers of samples. The shaded funnel is a symmetric parametric 95% confidence interval, based on 1000 replicate draws [44]. The dashed red line is the extrapolation out to the asymptotic minimum estimator (right solid circle) of 446 proteins [68]. Using the estimator of Chao *et al.* [44], an additional 18 MudPIT samples would be needed to reach this asymptote. Koziol *et al.* [46] describe other sampling models with these data.

Figure 3.

Proteomic sampling curve for the pitcher plant mosquito larva *Wyeomyia smithii*. The solid red line is a rarefaction curve illustrating the number of unique proteins identified as a function of the number of identified tryptic peptides. A total of 995 tryptic peptides were mapped to 98 unique proteins (closed circle). The rarefaction curve gives the expected number of proteins for a smaller random subsample of tryptic peptides. The shaded red funnel is a 95% confidence interval, based on 1000 replicates. The standard rarefaction analysis assumes that all tryptic peptides have an equal probability of detection. However, even if protein abundances were equal, proteins with high amino acid numbers would be over-represented in this analysis because they would be more likely to be cut into several tryptic peptides than proteins with low amino acid number. To adjust for these differences, the dashed blue line illustrates the expected rarefaction curve in which the probability of sampling a protein is proportional to $1/W$ where W is the number of amino acids it is comprised of. Amino acid numbers in this data set ranged from 217 (Nacalalpha-PB_FBgn0086904) to 8648 (bt-PC_FBgn0005666). This analysis suggests that protein diversity is higher than would be estimated from a standard rarefaction sampling curve. Data modified from Gotelli *et al.* [19].