Vol. 332: 247–248, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

THEME SECTION

Introducing genomics, proteomics and metabolomics in marine ecology

Idea and coordination: **Stewart C. Johnson, Howard I. Browman**

CONTENTS

Resale or republication not permitted without written consent of the publisher

Introduction

Stewart C. Johnson[1,*], Howard I. Browman[2]

[1]**Institute for Marine Biosciences, National Research Council, 1411 Oxford Street, Halifax, Nova Scotia B3H 3Z1, Canada**
[2]**Institute of Marine Research, Austevoll Research Station, 5392 Storebø, Norway**

Genomics, proteomics and metabolomics, used alone, in combination with each other and/or with more traditional methods, are fields of study that are rapidly transforming many areas of biological and biomedical research. They have enabled the transition from sequential studies of single genes, proteins or metabolites to what might be considered a more 'ecological approach', involving the simultaneous study of many components and their interactions with the environment (from pathways, through cell tissues to whole organisms and communites) (Hollywood et al. 2006, Joyce & Palsson 2006). The development of these fields has been supported by the concurrent development of many new technologies and methods such as quanitative PCR, RNA interference assays, and fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometry. Many of these technologies and methods are now used to address fundamental smaller-scale questions in areas such as ecology, biodiversity and evolution.

With the exception of the use of genomics to address questions about the diversity and ecology of marine microbial communities ('metagenomics', i.e. Venter et al. 2004, Delong et al. 2006, Sogin et al. 2006 and references therein), these fields have as yet not been broadly applied in marine ecology. The goal of this Theme Section (TS) is to provide an introduction to these fields, including information on how they have been applied (or could be applied) to address questions in marine ecology. Contributors were invited to explore questions such as: Is knowledge gained by the application of such technologies in marine ecology worth the money? Will technologies such as DNA bar coding ever replace traditional taxonomic studies? Have research areas such as environmental genomics

*Email: stewart.johnson@nrc-cnrc.gc.ca

© Inter-Research 2007 · www.int-res.com

met research expectations? What is the scientific value of large-scale genomic sequencing of marine animals? What factors limit the application of these technologies in the marine sciences? How can marine scientists be better trained to take advantage of such technologies? How can scientists with genomics, proteomics and metabolomics skills be encouraged to address questions in marine ecology?

Several common themes unite the contributions to this TS:

(1) Improved genomics resources (i.e. gene/protein sequences) for marine organisms will greatly facilitate the application of these fields to questions in marine ecology. To date, the development of genomics resources for marine organisms has been primarily focussed on marine microbes (see Thomas et al., in this TS). However, genomics resources for other taxa are at present limited to the full genome sequence of the 'model species', the purple sea urchin *Strongylocentrotus purpuratus*. Genomics sequencing efforts for other model and non-model species, including the diatom *Thalassiosira pseudonana*, the surf clam *Spisula solidissima*, the sea squirts *Ciona intestinalis* and *Ciona savignyi*, the tunicate *Oikopleura dioica*, the little skate *Leucoraja erinacea*, and the mollusc parasite *Perkinsus marinus*, are in progress. As noted by Dupont et al. (in this TS), both small- and large-scale expressed sequence tag (EST) resources and other genomic tools are becoming more available for species from a wide range of taxa. However, taken together, all of these resources still cover only a small fraction of marine taxa. With time, these resources will become more numerous as increased sequencing speed and reduced cost make genomic studies of marine organisms feasible for more research groups.

(2) Multidisciplinary teams, and the sharing of source materials and information, will add value to marine ecological research. Contributors to this TS emphasize the importance and value of developing multidisciplinary teams to plan, conduct, analyze and interpret the large amounts of information generated by these fields of study. However, in order to realize their full potential, it will be necessary to integrate these data with classical ecological approaches and knowledge. Distel (in this TS) discusses the importance of the preservation and sharing of biological source materials, and the information obtained from them. Providing research groups with methods to access samples and information that they would not normally be able to obtain will help to promote the multidisciplinary culture that is necessary to take full advantage of these fields of study when applied to marine ecological research.

(3) Data management, data sharing, other bioinformatics resources and knowledge are needed to extract meaningful biological information from large complex data sets. Contributors to this TS emphasize the fact that genomics, proteomics and metabolomics generate extremely large data sets that are difficult to interpret. The availability of bioinformatics resources, and personnel who are knowledgeable and skilled in their application, often limits the success of such studies. In fact, relatively simple processes such as data storage and data sharing often exceed the capacities of many research laboratories. Within each of these fields, the development of resources and tools with which to interpret data is evolving rapidly. In addition, new resources and tools to integrate 'omics' data sets, with the goal of understanding biology at the systems level, are also becoming more widely available (de Keersmaecker et al. 2006, Joyce & Palsson 2006). The ability to understand and appropriately utilize these bioinformatic resources and tools requires a great deal of training and is an area of expertise in itself. Future marine ecologists will need these skills and thus appropriate training at both the undergraduate and graduate level needs to be considered.

It is our hope that this TS will stimulate discussion within the marine ecological community, as well as encourage interactions between marine ecologists and other research groups that routinely use these fields of research. The goal is to develop the relationships and networks that would enable the formation of multidisciplinary teams that are so crucial for obtaining funding for large-scale marine ecological research programs that utilize and, more importantly, integrate these fields.

Acknowledgements. S.C.J.'s research is supported by the National Research Council of Canada and Genome Canada. H.I.B.'s research, and his editorial activity for MEPS, are supported by the Institute of Marine Research, Norway, The Research Council of Norway, and the Inter-Research Science Center.

LITERATURE CITED

de Keersmaecker SCJ, Thijs IMV, Vanderleyden J, Marchal K. (2006) Integration of omics data: how well does it work for bacteria? Mol Microbiol 62:1239–1250

DeLong EF, Preston CM, Mince T, Rich V and 8 others (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

Hollywood K, Brison DR, Goodacre R (2006) Metabolomics: current technologies and future trends. Proteomics 6: 4716–4723

Joyce AR, Palsson BØ (2006) The model organism as a system integrating 'omics' data sets. Nature Rev Mol Cell Biol 7: 198–210

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta, JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. Proc Natl Acad Sci USA 103:12115–12120

Venter JC, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Vol. 332: 249–255, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Genomics-enabled research in marine ecology: challenges, risks and pay-offs

Gretchen E. Hofmann*, Sean P. Place

Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106-9610, USA

ABSTRACT: Genomics-enabled applications are becoming increasingly common in conjunction with research in marine ecology. In this Theme Section, we review the success of cases where techniques used to profile gene expression have been used to gain new insight into 3 areas of research: symbioses in marine invertebrates, physiological responses to environmental conditions, and examining the determinants of species-range boundaries in marine ecosystems. In addition, we briefly discuss the challenges facing new practitioners of these techniques, including an overview of essential equipment to conduct research in ecological genomics.

KEY WORDS: DNA microarrays · Gene expression · Genomics · Marine ecology · Transcriptomics

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Numerous recent reviews have highlighted the increased use of genomics-based techniques to address central questions in ecology and evolution (Gibson 2002, Jackson et al. 2002, Feder & Mitchell-Olds 2003, Klaper & Thomas 2004, Ranz & Machado 2006, van Straalen & Roelofs 2006). Increasingly, these approaches are being applied in marine systems with outstanding results (Hofmann et al. 2005, Wilson et al. 2005). The goal of the present study is to highlight the costs and benefits of genomics-based approaches in marine ecology, concentrating on the practical aspects and the hurdles that are faced by a new practitioner of these molecular tools. In addition, we have strategically chosen examples and particular techniques that study the transcriptome (i.e. a collection of all mRNAs in a cell) as a means to focus the article, illustrate the power of this approach, and to highlight the advances made using techniques that profile patterns of gene expression (Schena et al. 1995, Gracey & Cossins 2003, Allison et al. 2006). It should be noted that there are numerous other examples within marine ecological genomics (Venter et al. 2004), and other important analytical tools (Marsh & Fielman 2005). Broadly focused discussions of genomics specific to ecological interests can be found in van Straalen et al. (2006) and recent reviews (Gibson 2002, Thomas & Klaper 2004).

IT'S WORTH THE EFFORT: INSIGHTS INTO BASIC AND APPLIED PROCESSES

For those considering entry into the genomics fray, a frequently asked question is: Are the costs of these technologies really worth it in proportion to what is being learned? In short, we believe the answer is an emphatic 'Yes'. The best way to illustrate this is by example. Thus, we have chosen studies of marine organisms that have used various methods (e.g. quantitative realtime PCR and cDNA [complementary DNA] macro- and microarrays) to profile gene expression in an ecological context (Appendix 1). Each of the examples share 2 characteristics: they each use a non-model marine organism, and they have each made leaps ahead in their respective fields owing to the insight generated by examining mRNA expression in their respective experimental systems. These examples include: (1) studying the mechanisms involved in important species interactions; in this case, in cnidarian-dinoflagellate symbiosis; (2) assessing the physiology of individuals such as the stress response; and

*Email: hofmann@lifesci.ucsb.edu

© Inter-Research 2007 · www.int-res.com

(3) examining expression across environmental gradients and linking physiological responses to large-scale ecological processes such as the determinants of biogeographic ranges in marine organisms.

Species interactions: exploring the cnidarian-algal symbioses

Some of the more compelling stories resulting from the use of DNA microarrays in marine ecology are the gene expression studies that identify genes involved in the mutualistic relationship between cnidarians and their intracellular algal symbionts. The outcome of this research illustrates the power of the 'discovery' process in genomics: results contribute to what is already known, and can often provide novel insight into important mechanisms. A recent study on the sea anemone *Anthopleura elegantissima* is exemplary of the utility of a genomics approach: Weis and colleagues demonstrated that the mechanisms involved in the maintenance and regulation of the relationship is perhaps more complex than might be expected (Rodriguez-Lanetty et al. 2006). In a comparison of symbiotic and aposymbiotic anemones, investigators found that genes from numerous metabolic processes displayed variation (Rodriguez-Lanetty et al. 2006). Using DNA microarray-based transcriptome analysis, 28 host genes were shown to vary in the symbiotic state; of these 28, functional-group analysis indicated that the results were underscoring that symbiosis had a more global effect on the host metabolism, rather than revealing a suite of genes unique to the symbiotic state (Rodriguez-Lanetty et al. 2006). In the supporting category, genes involved in lipid metabolism changed in a predictive fashion (i.e. some synthetic enzymes were down-regulated, and degradative ones were up-regulated). In the novel category, the study provided unprecedented insight into how apoptosis and cell-cycle genes may be related to maintaining the symbiosis by controlling the life of the host cell, something that investigators had observed in other symbioses, but was very new evidence in the cnidarian system.

Similarly powerful tools to assess the interaction of the host invertebrate and the algal symbiont are being built through ongoing efforts with coral genomics. Given the recent observations regarding changes in the strain of *Symbiodinium* that correlated with environmental conditions (Baker 2003, Rowan 2004), there is increasing evidence that the flexibility of the host–symbiont combination may be subject to environmental regulation. In order to characterize the nature of the symbiosis, Medina and colleagues have been constructing cDNA libraries for different stages of 2 important Caribbean coral species, *Acropora palmata* and *Monastraea faveolata*, and eventually will use these to assess gene expression that is linked to the symbiosis in stages ranging from eggs to adults in colonies (Schwarz et al. 2006). Similarly, a recent annotated cDNA library for the squid–*Vibrio* symbioses will facilitate research on this invertebrate–bacterial system (Chun et al. 2006). These efforts will certainly pay off enormously for investigators, providing new foundational data that can be used to form hypotheses about topics ranging from how the symbiosis is established to what modulators may be regulating the presence of the intracellular symbionts.

Organismal-level studies: individual performance and stress responses

Another emerging application for microarray-based transcript profiling is being found in studies that address organismal physiology. Here, suites of differentially regulated genes provide 'physiological fingerprints' of an organism's response to changes in abiotic conditions, especially with respect to stress. Successfully demonstrated in model organisms such as yeast (Gasch et al. 2000) and *Arabidopsis* (Seki et al. 2002, Rizhsky et al. 2004), this approach has been applied to marine organisms to assess response to short-duration changes in temperature (Gracey et al. 2001, Podrabsky & Somero 2004, Buckley et al. 2006), to disease (Dhar et al. 2003), and in studies that use organisms as biosensors in response to toxins (Klaper & Thomas 2004, Almeida et al. 2005, Dondero et al. 2006). Although the cost of transcriptomics has been called into question for the assessment of stress in some cases (Feder & Walser 2005), these techniques are appropriate for those seeking a deeper understanding of mechanisms because such studies can provide the foundations for future hypothesis testing. Especially for investigators interested in physiological or cellular responses, gene-expression profiling can be very informative and impart information about single gene families (Jenny et al. 2004), or provide insight into patterns of expression in specific biochemical pathways (Gracey et al. 2004, Buckley et al. 2006).

In addition to stress responses, microarray applications are being applied to basic questions such as how a particular genotype leads to the phenotype — i.e. inter-individual variation in natural populations (Oleksiak et al. 2002) — and, interestingly, how gene regulation contributes to this process (Ranz & Machado 2006). For example, researchers working on killifish have shown distinct differences in gene expression in individuals with varying performance abilities (Oleksiak et al. 2005). These data are intriguing in that they suggest that physiological performance is multifactor-

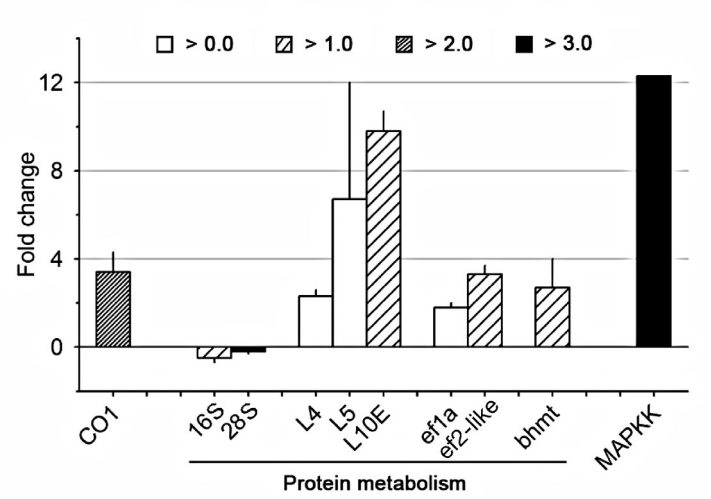

Fig. 1. Differences in gene expression detected by macroarray analysis of *Strongylocentrotus purpuratus* early, 4-arm echinoplutei following a 15 to 25°C transition (K. T. Fielman & G. E. Hofmann unpubl. results). Control (15°C) and treatment (25°C) larvae were held at their exposure temperatures for 24 h prior to RNA extraction. Changes in 3072 gene transcripts were evaluated using macroarrays and software provided by the Sea Urchin Genome Project. Fold change in gene expression as a function of temperature increase is shown on the *y*-axis for 10 genes: cytochrome oxidase 1 (CO1), mitochondrial ribosomal RNA (16S), cytosolic ribosomal RNA (28S), ribosomal proteins (L4, L5, L10E), translation elongation factors (ef1a and ef2-like), betaine-homocysteine methytransferase (bhmt), and a mitogen activated protein kinase kinase (MAPKK). Bars indicate the magnitude of change based on locally normalized data. This macroarray-based transcriptional profiling illustrates the response among larval gene groups important in aerobic respiration (CO1), protein metabolism, and cell signaling (MAPKK)

ial, the sum of perhaps subtle changes in numerous metabolic pathways. Furthermore, for the genomics practitioner, such differences among individual specimens must be considered in experimental designs.

Exploring environmental gradients: biogeographic range patterns in marine populations

A central aim in ecology is to determine the processes and mechanisms that set species biogeographic range boundaries (Gaston 2003), and genomics-enabled techniques are contributing to answers for these types of questions. Increasingly, marine ecologists are interested in understanding the physiological state of an organism across its range (Partensky et al. 1999, Somero 2005, Sorte & Hofmann 2005, Sagarin & Somero 2006, Stillman et al. 2006, Osovitz & Hofmann 2007). Genomic approaches, particularly the use of gene-expression profiling, have a great potential to help guide this discussion by providing insight into organismal performance across a variety of spatial scales.

Thus, perhaps one of the greatest utilities of genomic techniques is the illumination of the response to, and thus role of, temperature in effecting organismal distribution, an often complex response that is difficult to comprehensively quantify. Here, genomics-enabled techniques have been used successfully to examine 2 ecologically significant processes that contribute to species distribution: thermotolerance and dispersal. First, as a functional example, a study profiling gene expression in larvae of the purple sea urchin *Strongylocentrotus purpuratus* found that the expression of genes involved in protein metabolism and cell signaling was strongly affected by high temperature stress (Fig. 1). Results such as these contribute insight into dispersal recruitment processes in marine invertebrates, especially the presumptive role of temperature (Gaylord & Gaines 2000). Also in the dispersal category is the routine application of genomics approaches in microbial ecology (Zhou 2003). In the marine environment, using quantitative real time PCR (qPCR), oceanographers have identified variation in the distribution of *Prochlorococcus* spp. ecotypes that strongly correlated with temperature as a function of depth in the Atlantic Ocean (Johnson et al. 2006). Importantly, both of these examples highlight significant advances made by the application of genomic-scale tools to the assessment of overall patterns of gene expression.

LIMITATIONS AND CHALLENGES

Although divergent in focus, the examples of success given above all share a similar subtext: each project was faced with a series of challenges related to working on a non-model organism with emerging technology in a discipline to which genomic technology was largely foreign. These challenges are worth discussing in detail because the bar is moving in terms of how great these hurdles may be, and to what extent they can be ameliorated within marine ecology.

There are several challenges for the marine ecologist making the leap into functional genomics, each having a unique timeframe for resolution. Obviously, funding is a major hurdle. These techniques are expensive, requiring access to costly equipment, and involve the work of experienced researchers for whom salaries are required. The real-world consequence of this situation is that collaborations with colleagues that have expertise and resources in molecular biology are essential. One of the biggest initial decisions is how to partition the work: how much can your own group do, and what proportion of the work should be performed 'out of house'? In many cases, someone with minimal training in molecular biology can handle the basics of preparing cDNA libraries and DNA sequencing; in contrast,

printing and scanning arrays may require collaboration with the neuroscientist down the hall, in another department of even a separate academic or commercial institution.

Another major hurdle, and perhaps one of the more intractable ones, is access to equipment. Appendix 2 describes some of the basic equipment, reagents, and personnel associated with the methods of transcriptome analysis. No matter which method is employed, any laboratory undertaking a functional genomics project needs a fundamental set of equipment (e.g. thermal cycler, electrophoresis equipment, spectrophotometers) to perform basic molecular work in-house. While much of the smaller equipment purchases inevitably fall within an individual laboratory's budget, many departments have been successful in securing funds for core facilities that house and/or operate many of the specialty equipment items listed in Appendix 2. In addition, many processes that require the most expensive tools (sequencing, array construction, DNA library construction, etc.) can often be contracted out to genomic facilities for less than it would cost in-house when personnel, reagents, service contacts for major equipment, and time to successful completion are tallied. Ultimately, these techniques may prove too expensive for widespread use, i.e. the development of a comprehensive cDNA or EST library for all species of interest, but alternative techniques such as qPCR could still be applied (Osovitz & Hofmann 2005). Although restricted to a specific set of *a priori* selected genes, qPCR has a significant advantage over the highly complex methods mentioned previously. Since the construction of a comprehensive cDNA library for the study organism is not a prerequisite, equipment and reagent costs are comparable with routine molecular applications. However, research on non-model organisms remains significantly hindered by the lack of readily available sequence information.

Perhaps one of the most daunting hurdles for new investigators in the field of functional genomics is related to bioinformatics: namely, how to mine the voluminous raw data once it is in hand, and how to best analyze large data sets. Fortunately, the expense of this hurdle is not much more than the usual software bundle, and programs are often freely distributed via the Internet. Still, the learning curve associated with present-day bioinformatics analysis software can be steep, even with regard to the current trend of moving from platform- and programming language-specific command-line execution to platform-independent, user-friendly interfaces. Hundreds of analysis programs are available for array data alone, all of which will allow some degree of normalization, clustering, and hierarchical analysis of the raw data. For ecologists familiar with modeling algorithms, this may prove to be no hurdle at all because they may quickly adjust to the new, yet familiar parameters and analysis approaches.

IS THERE ANY GOOD NEWS?

In summary, and in our opinion, there is indeed good news in this arena. As described herein, exciting science is being conducted at the interface of genomics and marine ecology. Owing to the redundancy of the basic molecular biology techniques, many young investigators versed in molecular ecology have the tools with which to conduct these experiments. In addition, from a resource perspective, the printing of microarrays is becoming more accessible as more central facilities acquire the printers and microarray scanners. Individual principal investigators are also finding that prices for these major pieces of equipment are reducing and are affordable for a single laboratory group.

Cross-species hybridizations are increasingly commonplace and seem to be yielding reliable results (Ji et al. 2004, Renn et al. 2004). Although initially received with much skepticism, cross-species hybridizations have begun to gain acceptance within the comparative community. The feasibility of obtaining biologically meaningful data has started to be systematically assessed with favorable results. For example, using a microarray derived from the cDNA library of a cichlid fish, Renn et al. (2004) conducted heterologous hybridizations with several divergent fish species, showing that consistent expression profiles can be achieved for species that diverged as long as 65 million years ago. Combined with the efforts of genome projects currently underway for species central to ecological and evolutionary studies, these results show great promise for the application of molecular-based approaches to the elucidation of complex phenotypes.

Overall, these findings indicate that continued cooperation among colleagues within these fields will facilitate the use of genomic approaches. Increased cooperativity is already apparent within the marine and aquatic biology communities. Consortia such as the Marine Genomics group at the University of South Carolina (McKillen et al. 2005) and the consortium for Genomics Research on All Salmon (GRASP) exemplify the teamwork, cooperativity, and resource sharing that will ensure continued success in marine ecological and environmental genomics. Also, groups interested in genomics in specific biogeographic regions (Clark et al. 2004, Schwarz et al. 2006) are moving towards important shared resources for the marine community. Finally, governmental support for marine genomics is rapidly emerging. For example, in the USA, the Joint

Genome Institute (JGI) has assisted with sequencing large numbers of genes from non-model marine organisms (e.g. coral, intertidal mussels, and crabs). Yet, strong funding support for young, interdisciplinary investigators who will train the next generation of researchers in marine genomics is warranted.

Acknowledgements. We thank Dr. Kevin Fielman and Dr. Anne Todgham for helpful conversations and editorial comments that improved the manuscript. We acknowledge the US National Science Foundation for financial support during the course of writing (NSF grants OCE-0425107 and ANT-0440799 to GEH). This is Contribution No. 225 from PISCO (the Partnership for Interdisciplinary Studies of Coastal Oceans) funded primarily by the Gordon and Betty Moore Foundation and David and Lucile Packard Foundation.

LITERATURE CITED

Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nature Rev Genet 7:55

Almeida JS, McKillen DJ, Chen YA, Gross PS, Chapman RW, Warr G (2005) Design and calibration of microarrays as universal transcriptomic environmental biosensors. Comp Funct Genom 6:132–137

Baker AC (2003) Flexibility and specificity in coral-algal symbiosis: diversity, ecology, and biogeography of *Symbiodinium*. Annu Rev Ecol Evol Syst 34:661–689

Buckley BA, Gracey AY, Somero GN (2006) The cellular response to heat stress in the goby *Gillichthys mirabilis*: a cDNA microarray and protein-level analysis. J Exp Biol 209:2660–2677

Chun CK, Scheetz TE, Bonaldo MF, Brown B and 20 others (2006) An annotated cDNA library of juvenile *Eupryma scolopes* with and without colonization by the symbiont *Vibrio fischeri*. BioMed Central (BMC) Genom 7:154

Clark MS, Clarke A, Cockell CS (2004) Antarctic genomics. Comp Funct Genom 5:230–238

Dhar AK, Dettori A, Roux MM, Klimpel KR, Read B (2003) Identification of differentially expressed genes in shrimp (*Penaeus stylirostris*) infected with white spot syndrome virus by cDNA microarrays. Arch Virol 148:2381–2396

Dondero F, Piacentini L, Marsano F, Rebelo M, Vergani L, Venier P, Viarengo A (2006) Gene transcription profiling in pollutant exposed mussels (*Mytilus* spp.) using a new low-density oligonucleotide microarray. Gene 376:24

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nature Rev Genet 4:649–655

Feder ME, Walser JC (2005) The biological limitations of transcriptomics in elucidating stress and stress responses. J Evol Biol 18:901–910

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241–4257

Gaston KJ (2003) The structure and dynamics of geographic ranges. Oxford University Press, Oxford

Gaylord B, Gaines SD (2000) Temperature or transport? Range limits in marine species mediated solely by flow. Am Nat 155:769–789

Gibson G (2002) Microarrays in ecology and evolution: a preview. Mol Ecol 11:17–24

Gracey AY, Cossins AR (2003) Application of microarray technology in environmental and comparative physiology. Annu Rev Physiol 65:231–259

Gracey AY, Troll JV, Somero GN (2001) Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis*. Proc Natl Acad Sci USA 98:1993–1998

Gracey AY, Fraser EJ, Li W, Fang Y, Taylor RR, Rogers J, Brass A, Cossins AR (2004) Coping with cold: an integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. Proc Natl Acad Sci USA 101:16970–16975

Hofmann GE, Burnaford JL, Fielman KT (2005) Genomics-fueled approaches to current challenges in marine ecology. Trends Ecol Evol 20:305–311

Jackson RB, Linder CR, Lynch M, Purugganan M, Somerville S, Thayer SS (2002) Linking molecular insight and ecological research. Trends Ecol Evol 17:409–414

Jenny MJ, Ringwood AH, Schey K, Warr GW, Chapman RW (2004) Diversity of metallothioneins in the American oyster, *Crassostrea virginica*, revealed by transcriptomic and proteomic approaches. Eur J Biochem 271:1702–1712

Ji W, Zhou W, Gregg K, Yu N, Davis S, Davis S (2004) A method for cross-species gene expression analysis with high-density oligonucleotide arrays. Nucleic Acids Res 32:e93, doi:10.1093/nar/gnh084

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311:1737–1740

Klaper R, Thomas MA (2004) At the crossroads of genomics and ecology: the promise of a canary on a chip. BioScience 54:403–412

Marsh AG, Fielman KT (2005) Transcriptome profiling of individual larvae of two different developmental modes in the poecilogonous polychaete *Streblospio benedicti* (Spionidae). J Exp Zool 304B:238–249

McKillen D, Chen Y, Chen C, Jenny M and 7 others (2005) Marine genomics: a clearing-house for genomic and transcriptomic data of marine organisms. BioMed Central (BMC) Genom 6:34

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32:261

Oleksiak MF, Roach JL, Crawford DL (2005) Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. Nat Genet 37:67

Osovitz CJ, Hofmann GE (2005) Thermal history-dependent expression of the *hsp70* gene in purple sea urchins: biogeographic patterns and the effect of temperature acclimation. J Exp Mar Biol Ecol 327:134

Osovitz CJ, Hofmann GE (2007) Marine macrophysiology: studying physiological variation across large spatial scales in marine systems. Comp Biochem Physiol

Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63:106–127

Podrabsky JE, Somero GN (2004) Changes in gene expression associated with acclimation to constant temperatures and fluctuating daily temperatures in an annual killifish *Austrofundulus limnaeus*. J Exp Biol 207:2237–2254

Ranz JM, Machado CA (2006) Uncovering evolutionary patterns of gene expression using microarrays. Trends Ecol Evol 21:29–37

Renn S, Aubin-Horth N, Hofmann H (2004) Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BioMed Central (BMC) Genom 5:42

Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defense pathways collide. The

response of *Arabidopsis* to a combination of drought and heat stress. Plant Physiol 134:1683–1696

Rodriguez-Lanetty M, Phillips W, Weis V (2006) Transcriptome analysis of a cnidarian — dinoflagellate mutualism reveals complex modulation of host gene expression. BioMed Central (BMC) Genom 7:23

Rowan R (2004) Coral bleaching: thermal adaptation in reef coral symbionts. Nature 430:742

Sagarin RD, Somero GN (2006) Complex patterns of expression of heat-shock protein 70 across the southern biogeographical ranges of the intertidal mussel *Mytilus californianus* and snail *Nucella ostrina*. J Biogeogr 33:622–630

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Schwarz J, Brokstein P, Manohar C, Coffroth MA, Szmant A, Medina M (2006) Coral reef genomics: developing tools for functional genomics of coral symbiosis. In: Proc 10th Int Coral Reef Symp. Japanese Coral Reef Society, Okinawa, p 274–281

Seki M, Narusaka M, Ishida J, Nanjo T and 14 others (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 31:279–292

Somero G (2005) Linking biogeography to physiology: evolutionary and acclimatory adjustments of thermal limits. Front Zool 2:1

Sorte CJB, Hofmann GE (2005) Thermotolerance and heat-shock protein expression in Northeastern Pacific *Nucella* species with different biogeographical ranges. Mar Biol 146:985–993

Stillman JH, Teranishi KS, Tagmount A, Lindquist EA, Brokstein PB (2006) Construction and characterization of EST libraries from the porcelain crab, *Petrolisthes cinctipes*. Integr Comp Biol 46:919–930

Thomas MA, Klaper R (2004) Genomics for the ecological toolbox. Trends Ecol Evol 19:439–445

van Straalen NM, Roelofs D (2006) Introduction to ecological genomics. Oxford University Press, Oxford

Venter JC, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Wilson K, Thorndyke M, Nilsen F, Rogers A, Martinez P (2005) Marine systems: moving into the genomics era. PSZN I: Mar Ecol 26:3–16

Zhou J (2003) Microarrays for bacterial detection and microbial community analysis. Curr Opin Microbiol 6: 288

Appendix 1. Techniques used in analysis of the transcriptome

With the continuous refining of various methods used to profile gene expression, these techniques are becoming applied to non-model systems with ever greater numbers. Below we describe the more commonly used methods for analyzing the transcriptome

Quantitative real-time PCR (qPCR): a modification of the PCR in which cDNA is quantified after each round of amplification (real-time) as opposed to the end-point analysis of standard PCR reactions. Via reverse transcription, qPCR is used to quantify low abundance messenger RNA (mRNA), enabling a researcher to quantify relative gene expression at a particular time, or in a particular cell or tissue type. Two common methods of quantification involve the use of fluorescent dyes, which intercalate with double-stranded DNA, or modified DNA oligonucleotide probes that fluoresce when hybridized with complementary DNA. The amount of fluorescence emitted by these dyes is directly proportional to the number of amplicons produced in the reaction, and is measured by an optical module within a thermal cycler. Through the use of multiple dye combinations, researchers can 'multiplex' (monitor multiple genes in a single reaction); however, this technique requires a much greater understanding of the chemistry involved as well as more advanced real-time systems

Macroarray: a collection of DNA sequences, tens to thousands of bp in length, spotted onto a reusable membrane (generally nylon). With a spot size of 300 µm, these arrays typically hold hundreds to thousands of features on a single membrane. Hybridization of the array with samples labeled with a fluorescent or radiolabeled reporter offers the simultaneous analysis of relative sequence abundance for thousands of genes in a sample. While offering the advantage of being reusable several times, analysis of the filters requires the use of a fluorescent scanner (e.g. Storm Phosphorimager, Molecular Dynamics) or the use of relabeled probes. In addition, owing to the size of the features spotted on the array, several membranes may be required to analyze the entire genome of an organism, entailing that the researcher generate more sample for each analysis

Microarray: similar to macroarrays, microarrays are a collection of single DNA sequences immobilized on a solid surface, in this case glass. Utilizing features of ≤200 µm in diameter, microarrays are printed in much greater densities, often containing tens to hundreds of thousands of genes on a single chip. By hybridizing microarrays with 2 alternatively labeled samples, researchers can measure the relative abundance of the entire transcriptome in a given sample. While allowing the entire genome to be printed on a single chip, these arrays are not reusable and require costly equipment such as robotic arrayers for printing and dual laser scanners for fluorescence capturing

Appendix 2. Basic equipment used in genomics research

Thermal cycler: thermal cyclers are used for PCR amplification of a specific DNA template. They use temperature-controlled blocks to cycle between programmed periods of DNA denaturation, primer annealing, and sequence elongation. Each 'cycle' of amplification results in exponential increases in the pool of the DNA sequence of interest. This particular piece of equipment is of use to any molecular application and is essential for the amplification of cDNA in preparation for microarray/macroarray printing as well as downstream sequencing applications

Real-time system: this system combines thermal cycling, fluorescence detection, and application-specific software to provide an integrated platform for the detection and quantification of nucleic acid sequences. These systems can become quite complicated with automated components for high-throughput applications; however, a standard 96-well compatible system will suffice for the most advanced ecological genomics laboratory. Reagents for real-time PCR can be a significant cost, requiring the regular purchase of fluorescent dyes, reverse transcriptase enzyme, *Taq* polymerase enzyme, and high purity oligonucleotide primers. The thermal cycler contained within these systems can often suffice for general thermal cycling applications as well

UV-spectrophotometer: used for the quantification of DNA/RNA concentration as well as the purity of a sample based on 260:280 nm ratios. These instruments range in price largely based on the number of samples (single samples up to 384 well plates) and the volume of sample used (1 µl up to 5 ml)

Horizontal gel electrophoresis: used for separation of DNA/RNA sequences, via agarose gels, based on fragment size for visualization, non-quantitative determination of concentration, molecular integrity, and even purification of fragments

cDNA library: a cDNA library refers to a complete, or nearly complete, set of all mRNAs contained within a cell or organism. Researchers use an enzyme called reverse transcriptase, which produces a DNA copy (cDNA) of each mRNA strand. These cDNAs are collectively known as the 'library'. Production of the library is a lengthy and often technically challenging endeavor. If not collaborating with a laboratory familiar with this process, all is not lost. Recently, several independent companies have begun to create custom libraries, and a few even specialize in non-model organisms. The price for the custom production of a library has come down in recent years and one can expect to pay between US $5000 and $10 000, depending on the company and if you have the companies pick individual clone sets and print the clones on membranes in addition to producing the library

Dual laser microarray scanner: a confocal laser scanning device used to detect and quantify hybridization signals (532 and 635 nm), and specifically designed to scan DNA microarrays fabricated on glass slides using Cy3 and Cy5 fluorescent labels

Microarrayer: an environmentally controlled robotic printer capable of printing biological samples on standard-sized glass slides. Printing is from microtiter plates with 96 or 384 wells. Features are printed as 100 to 200 µm spots, size depending on tip type used; these robots can array hundreds of thousands of features per slide

DNA sequencer: automated processor used for determining the exact order of the bases A, T, C and G in a piece of DNA. The most commonly used method of sequencing DNA—the dideoxy or chain termination method is achieved by including in each reaction a nucleotide analogue that cannot be extended and thus acts as a chain terminator. In essence, DNA is used as a template to generate a set of fragments that differ in length from each other by a single base. Fragments are then separated by size, and the bases at the end are identified, recreating the original sequence of the DNA. The primers or nucleotides included in the reactions contain different fluorescent labels, allowing DNA strands terminating at each of the 4 bases to be identified. Reaction products are separated by gel electrophoresis. As the DNA strands passes a specific point, the fluorescent signal is detected and the base identified. Many outside sequencing facilities are available to researchers, and are often more cost effective when high-throughput (i.e library annotation) sequencing is not necessary; however, time can be lost due to sample delivery to these facilities

Bioinformatics: basic computing capabilities are necessary for all the methods described in the present study. Most data from these processes are produced in digital formats and therefore require software packs compatible with the output from each specific approach taken

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: August 21, 2006; Accepted: September 21, 2006
Proofs received from author(s): February 8, 2007

Vol. 332: 257–273, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Marine ecological genomics: when genomics meets marine ecology

Samuel Dupont*, Karen Wilson, Mathias Obst, Helen Sköld, Hiroaki Nakano, Michael C. Thorndyke

Kristineberg Marine Station, 566 Kristineberg, 45034 Fiskebäckskil, Sweden

ABSTRACT: Genomics, proteomics and metabolomics (the 'omic' technologies) have revolutionized the way we work and are able to think about working, and have opened up hitherto unimagined opportunities in all research fields. In marine ecology, while 'standard' molecular and genetic approaches are well known, the newer technologies are taking longer to make an impact. In this review we explore the potential and promise offered by genomics, genome technologies, expressed sequence tag (EST) collections, microarrays, proteomics and bar coding for modern marine ecology. Methods are succinctly presented with both benefits and limitations discussed. Through examples from the literature, we show how these tools can be used to answer fundamental ecological questions, e.g. 'what is the relationship between community structure and ecological function in ecosystems?'; 'how can a species and the phylogenetic relationship between taxa be identified?'; 'what are the factors responsible for the limits of the ecological niche?'; or 'what explains the variations in life-history patterns among species?' The impact of ecological ideas and concepts on genomic science is also discussed.

KEY WORDS: Sequencing · ESTs · Microarrays · Proteomics · Barcoding

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Genome-based technologies are revolutionizing our understanding of biology at all levels, from genes to ecosystems. Genomics is the study of the genomes of various organisms in their entirety, while genetics tends to study genes individually or in linked groups, relating DNA sequences to proteins and ultimately to heritable traits (Van Straalen & Roelofs 2006). The term 'genomics' appeared in the 1980s as the name of a new journal (McKusick & Ruddle 1987), but the genomics revolution really began in 1990 with the Human Genome Project and since then, thanks to rapid developments in molecular biology technologies, genomics-based discovery has grown exponentially. For example, the new sequencing system developed by Margulies et al. (2005) will be capable of sequencing 25 million bases in a 4 h-period — about 100 times faster than current state-of-the-art systems — with the same reliability and accuracy. The genomes of more than 300 organisms have been sequenced and analyzed since the publication of the first complete genome in 1995, and today a new organism is sequenced nearly every week (Rogers & Venter 2005, Van Straalen & Roelofs 2006). The current challenge is no longer to collect sequence information but rather to analyze the data. Genomic approaches combine molecular biology with computing sciences, statistics and management. The intellectual infrastructure in genomics must be extended into bioinformatics (data storage and data query), computational biology (more complex, often hypothesis-driven analyses that may require the development of new algorithms and tools), and information technologies to share software and data.

Molecular ecology is a relatively new field in which techniques such as Polymerase Chain Reaction (PCR) and genetic engineering (recombinant DNA technology) has had an increasing role in the integration of genetic data with historical or field observations (White 1996). Through the study of single or small sub-

*Email: samuel.dupont@kmf.gu.se

© Inter-Research 2007 · www.int-res.com

sets of genes or small genomic regions (e.g. microsatellites), molecular ecology has been used to address classic questions in the areas of diversity, populations, and taxonomy. In contrast, the emerging field of ecological genomics is trying to answer larger ecological questions in areas such as nutrient cycling, population structure, life-history variations, trophic interactions, stress responses and ecological niches. Ecological genomics can be defined as 'the scientific discipline that studies the structure and functioning of a genome with the aim of understanding the relationship between the organism and its biotic and abiotic environments' (Van Straalen & Roelofs 2006). This new field crosses and interacts extensively with other disciplines such as microbiology, physiology, genetics and evolutionary biology. Ecological genomics investigates different levels of integration from the lower (functional mechanisms: physiology, biochemistry, cell biology, neuroscience, developmental biology etc.) to higher (ecology, evolution). The inclusion of 'function' is critical because the goal is to understand what genes/genomes and their variants do at higher levels of integration.

Marine ecological genomics is, then, the application of genomic sciences to attempt to understand the structure and function of marine ecosystems. Genomics provides biological information that is unobtainable by any other means, for example the biological capacities of marine organisms that underlie the ecology of oceanic ecosystems (see 'Genome sequencing: applications'). Approaches can include (1) whole genome sequencing of key organisms (e.g. genome comparison for phylogeny), or (2) genomic analysis of natural communities to understand how biodiversity supports ecosystem function (e.g. genomic analysis of microbial communities *in situ* with the concept of 'genome ecology', the collective genome in a given environment, also conceived as 'metagenomics'). For example, these approaches can be used to investigate life-history patterns (population ecology) and stress responses (physiological ecology).

Marine ecological genomics is a good example of a 21st century science that requires the mixing of scientific disciplines, hitherto historically and traditionally separated. Forging the link between marine ecologists, molecular biologists and genomics/bioinformatics scientists, and finding a common language, is a huge social challenge. The need for marine ecological genomics to be interdisciplinary is brought about by a number of factors including (1) the requirement for the use of highly specialized technologies, (2) the necessity for the development of new tools in key areas such as statistics and computational sciences, and (3) the lack of adequate funding for large-scale genome science research, especially in individual laboratories. As a consequence, this field is not always fully amenable to the individual or individual research group, and it is often essential and more strategically viable to develop coordinated networks of collaborative interdisciplinary laboratories. Nevertheless, some techniques are more affordable than others. For example, expressed sequence tag (EST) libraries and microarrays limited to genes associated with a specific function, tissue or response pathway can be manufactured at relatively low costs for small research groups (Held et al. 2004).

With the exception of microbial ecology, genomic studies have until recently only been performed on a rather limited number of classic model species such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus* or *Arabidopsis thaliana*. However, this is now changing and the number of new genomes is increasing (see www.ncbi.nlm.nih.gov, www.hgsc.bcm.tmc.edu/projects, http://genome.jgi-psf.org). The choice of the ideal model species for genomics is based on many practical (established reputation, genome size, possibility of genetic manipulation etc.) and scientific criteria (medical, biotechnological, agricultural or ecological significance, evolutionary position, comparative purpose, laboratory expertise etc.; see Feder & Mitchell-Olds 2003). This approach is not the tradition in ecology and there is a discrepancy between the available genomic models and ecologically interesting species. For example, *D. melanogaster* or *A. thaliana* are not sufficiently widespread in the environment and not very suitable for ecological studies. Moreover, no model is able to answer all questions. Consequently, the genomics revolution is the perfect time to move away from our fascination with model species, and the sequencing of the genomes of species such as *Ciona intestinalis* (http://ghost.zool.kyoto-u.ac.jp/indexr1.html, http://genome.jgi-psf.org/Cioin2/Cioin2.home.html) or the sea urchin *Strongylocentrotus purpuratus* (Sea Urchin Genome Consortium 2006, see also http://sugp.caltech.edu) as well as the amphioxus *Branchiostoma floridae* and the anemone *Nematostella* (www.jgi.doe.gov/sequencing/DOEprojseqplans.html) is the first step in this direction.

The marine ecology community must be prepared for the genomic era. The aim of this review is thus to explain general principles of the main genomic technologies and their applications to marine ecology with examples from the literature (for a more exhaustive presentation of genomic methods, see Van Straalen & Roelofs 2006). Genomic methods are succinctly presented with their strengths and limitations, and linked to marine ecological questions (Fig. 1). Marine ecological genomics is a new discipline merging genomics and marine ecology leading to new questions independent of both fields. Genomics is more than a toolbox

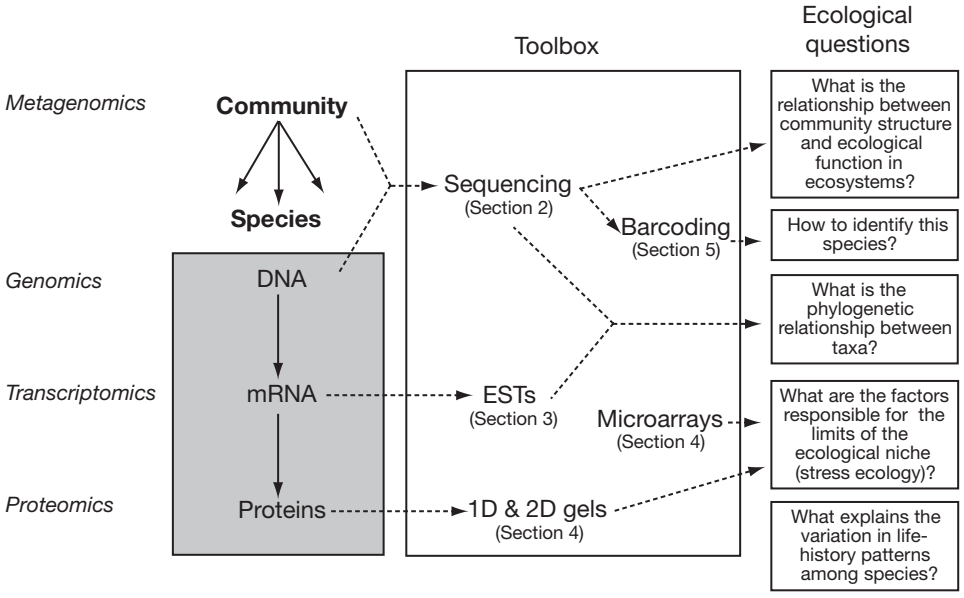

Fig. 1. Links between genomic tools and ecological questions

added to marine ecology; in the conclusion, therefore, some examples of feedback from ecology to genomics will be presented.

GENOME SEQUENCING

Method

Large-scale sequencing and first annotation is usually automated and based on the method initially developed by Sanger et al. (1977) and Smith et al. (1986). Nowadays, whole-genome sequencing is usually contracted out to commercial sequencing centres or is organised in collaborative networks comprising many different laboratories, often funded by national or international consortia (Van Straalen & Roelofs 2006). A complete description of methods and approaches is outside the scope of this article. For a list of websites and sequencing initiatives see 'Discussion'.

Applications

Environmental genome—microbial ecology

One of the most fundamental questions in community ecology is: what is the relationship between ecosystem processes and biodiversity? In other words, 'what do species do in ecosystems?' (Lawton 1994). In order to understand how biodiversity supports ecosystem function, it is necessary to estimate species diversity (richness, biomass, dominance structure, feeding groups) and functions (production, respiration, degradation of organic matter, nitrification etc.). This is particularly difficult in marine microbial communities where it is not always clear what constitutes a microbial species and it is only possible to characterize species that can be cultured. Here, genomics provides a solution to the problem by reconstructing diversity and functions from the environmental genome (partial or whole sequence from 'environmental samples', i.e. DNA extracted from a seawater sample). The DNA of all species in a microbial environment can be assembled and functions characterized without attempting to put them into culture or separate them according to species. For example, the genome of the anammox bacterium *Kuenenia stuttgartiensis* was recently deduced from the DNA sequenced from a whole microbial community (Strous et al. 2006). This will enable insight into the metabolism and evolution of this bacterium, which is responsible for removing up to 50% of fixed nitrogen from the ocean. Using similar approaches, it is also possible to compare two communities, detect functional genes indicative of key steps in cycles (nitrogen, sulphur etc.) or reconstruct functions without the need for culture (Van Straalen & Roelofs 2006).

With an estimated 2 million species of bacteria in pelagic zones, a density of billions of cells per litre and a richness of 163 species per millilitre of ocean water (Curtis et al. 2002, DeLong & Karl 2005), microbes are major players in the structure and dynamics of marine ecosystems. It is of crucial importance to understand microbial roles in oceanic primary production, global carbon cycling and functioning of the biosphere.

Unfortunately, in the oceans, most microbes (>99%) resist efforts to grow them in pure culture. In consequence, very little is known about their physiology and their role in the environment. These organisms can be categorized into phylotypes using rRNA genes amplified from environmental DNA extracts; however, this does not reveal the physiology, biochemistry or ecological function of uncultured microbes (Giovannoni & Stingl 2005). Ecological genomics appears to be a new culture-independent tool with which to analyze microbial community structure and function in natural and engineered environments. Microbial communities can be explored by isolating large fragments of DNA directly from the environment, sequencing the fragments and assigning function to the genes based on their similarity to known genes or on functional studies. This process is referred to as community genomics or metagenomics. The recent genomic survey of the Sargasso Sea microbial assemblage is a perfect example. This led to 1.6 billion bp of genome sequence information and about 1.2 million genes identified from the collective microbial assemblage (Tringe et al. 2005). Such data frequently leads to the discovery of new genes (e.g. photorhodopsin, Béjà et al. 2000), gene functions, novel metabolic pathways, and other previously unknown properties of micro-organisms. These data can also shed light on physiological properties and ecological functions without consideration of species. Using such an approach, it is also possible to identify the genes and biochemical pathways that differentiate species living in different environments.

Microbial genomes are relatively small and allow rapid and relatively inexpensive sequence determination (Béjà 2004, Steele & Streit 2005). Cyanobacteria are a good example. They are amongst the most widespread and relevant organisms in marine habitats, and the genus *Prochlorococcus* has a key role in terms of global primary production (Hess 2004). The observation of the absence of the nitrate reductase gene from the *Prochlorococcus* genome changed the way we think about the ecological role of this organism in pelagic systems (García-Fernàndez et al. 2004). Genome sequencing of several biodegradation-relevant micro-organisms has provided the first whole-genome insights into the genetic background of the metabolic capability and biodegradative versatility of these organisms (Pieper et al. 2004).

Comprehensive approaches to describe and interpret oceanic microbial diversity and processes are only now emerging. Genomics applied to microbial ecology is significantly expanding our understanding of marine microbial evolution, metabolism and ecology. This new technology is revealing the links between evolutionary, ecological and biogeochemical processes in natural marine microbial communities (DeLong & Karl 2005). Genomics applied to microbial ecology is a striking example of true and successful marine ecological genomics that enhances our understanding of the living marine system, and that will lead to a new generation of more realistic oceanographic simulations, including improved climate change projections (Doney et al. 2004).

Comparing genomes—phylogenomics

Genomics has changed the way we define the term 'species'. Whole genome comparisons (size, G/C content, number of genes, gene distribution, sequence etc.) allow identification of core similarities and differences at each level of complexity. Whole genome comparisons for different strains suggest that polymorphism is common and in some cases reflects adaptability to different habitats. The genomic era is now providing the opportunity for phylogenetics to resolve a number of outstanding evolutionary questions through an increase of resolving power (Delsuc et al. 2005).

Despite extensive research, high-level phylogenetic relationships amongst animals remain contentious. Studies have been based upon several developmental, morphological and, more recently, molecular tools. Two main hypotheses are proposed (Fig. 2): (1) the Acoelomata-Pseudocoelomata-Coelomata (A-P-C) hypothesis, supported by morphological and whole-genome studies, divides animals according to the presence/absence of a coelom, lined (or not) by mesoderm; (2) the Lophotrochozoa-Ecdysozoa-Deuterostomia (L-E-D) hypothesis, supported by genetic studies, divides animals into Protostomia-Deuterostomia based

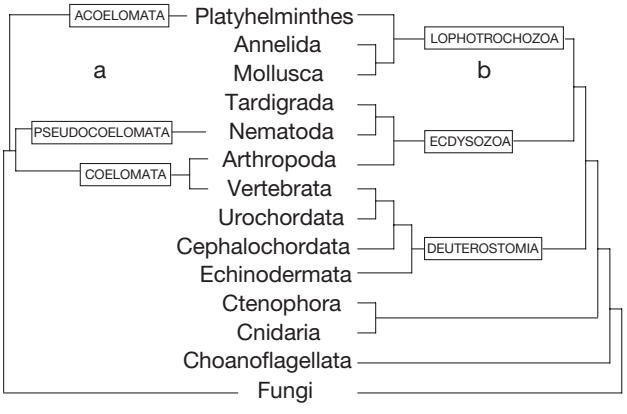

Fig. 2. Two main hypotheses of the relationship between animal phyla: the (a) Acoelomata-Pseudocoelomata-Coelomata phylogeny (A-P-C hypothesis) and (b) Lophotrochozoa-Ecdysozoa-Deuterostomia phylogeny (L-E-D hypothesis). From Jones & Blaxter (2005)

on origin of the mouth during development and the Lophotrochozoa-Ecdysozoa division of Protostomia based on moulting and presence of a lophophore. The use of molecular tools has resulted in some radical rearrangements of animal phyla. For example, phylogenetic analysis of 18S ribosomal DNA sequences supported the idea of Ecdysozoa (Winnepennincks et al. 1995, Aguinaldo et al. 1997, Adoutte et al. 1999, Peterson & Eernisse 2001). Nevertheless, analyses of whole genome sequences from a few species support older views (e.g. on human, fly, nematode and yeast by Mushegian et al. 1998, Blair et al. 2002, Wolf et al. 2004). These multigene analyses covered rather few taxa, and it is well known that the number of species represented in a phylogenic study can induce systematic artefacts on tree reconstruction. For example, genome-scale analyses are especially sensitive to long-branch attraction (Felsenstein 1978). For these studies, the usual outgroup is yeast, very distantly related to animals and species such as the nematode *Caenorhabditis elegans*. Here, nematodes move to the base of the tree, generating support for the A-P-C hypothesis (Mushegian et al. 1998, Blair et al. 2002, Wolf et al. 2004). When analyses are set up to avoid long branch attraction, they do not support A-P-C hypothesis but rather L-E-D; for example, analyses of rare insertions and deletions of genomic features in some animal genomes (Roy & Gilbert 2005) or analysis of data from ESTs (see 'Expressed Sequence Tags' below) in addition to complete genome sequences (Philippe et al. 2005). This last study demonstrated that if only yeast is used as an outgroup, nematodes emerge at the base of the tree. However, by using outgroups closer to animals, nematodes cluster close to arthropods as predicted by the L-E-D hypothesis. This clustering is also improved when biased genes (those with greatest evolutionary rate in some species) are removed. Only 12 of the 35 animal phyla are currently represented in genomic studies and the use of genomics in phylogeny is still at its infancy. More genomes need to be sequenced and new analytical tools (e.g. algorithms and software) should be developed.

EXPRESSED SEQUENCE TAGS (EST)

Method

Development of an EST library is often the first step when starting a genomic project on a novel organism. Complete genome sequencing provides information about genome organization and promoter regions etc. It is, however, a major investment and not likely to be applied to the majority of organisms that are subjects for scientific investigation. In contrast, genes of an increasing number of species are being investigated through generation of ESTs (www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). ESTs are cost-effective and provide a rapid strategy with which to identify genes of the investigated organism. The sequence information contributes to the understanding of the dynamics of genome expression patterns and thereby to the understanding of the biology of the organism. ESTs can be used in expression profiling, evolutionary and taxonomy studies, systematics etc. It is important to remember that these are expressed sequences (RNA) and will vary according to the temporal and spatial (tissue/organ) origin of the cDNA.

ESTs are usually obtained by sequencing clones from a cDNA library and can be assembled into an EST database containing the fragments of the sequenced cDNAs. The cDNA library can be made by the individual researcher or commercially by various companies that offer such services. In most commercial kits for library construction, it is possible to make several libraries using the same kit. Commonly, about 1 g of tissue or 1 mg of total RNA is used for a standard-sized library, a factor important to bear in mind if working with limited amounts of material. A subtraction library can be made by removing identical genes present in 2 libraries from different conditions, and here you produce a library containing genes differentially expressed according to the 2 conditions chosen (for example 2 temperatures, pH etc.). To produce an ordinary but enriched library, it is possible to use organisms from particular environmental conditions in order to enrich transcripts induced by that particular treatment. For example, Kore-eda et al. (2004) analysed the profile of differentially expressed genes of well-watered and salinity-stressed specimens from the common ice plant *Mesembryanthemum crystallinum*. The same number (2782) of ESTs from each library (total = 8346 ESTs) were randomly selected and analysed. Their result showed differential expression of known genes related to stress responses, and also of novel and/or functionally unknown genes that may have a novel role in the salinity stress response. A similar approach using a subtractive hybridisation library has been used successfully to analyse hierarchical behaviour in rainbow trout (Sneddon et al. 2005). A fascinating use of this comparative approach is evident from the work of Kuo et al. (2004), who constructed 2 complementary DNA (cDNA) libraries from RNA isolated from symbiotic and aposymbiotic *Aiptasia pulchella* in order to understand algal-cnidarian interactions. Their systematic analysis of these ESTs provides a useful database containing numerous putative candidate genes for further investigations.

Functional annotation of ESTs from ordinary cDNA libraries by basic local alignment search tool (BLAST)

comparisons commonly identify unique sequences that share significant similarities to nucleotide or amino acid sequences of genes with known as well as unknown functions. In addition, relatively large numbers of ESTs often do not significantly match any genes in public databases. These may represent previously unidentified genes. Typically, a subsequent clustering analysis will further reveal higher expression of ribosomal genes and genes coding for metabolic pathway proteins, structural proteins, cell cycle proteins and proteins involved in cellular defence and stress responses (Ogasawara et al. 2002, Hackett et al. 2005, Watanabe et al. 2005, Simon et al. 2006). Genes involved in such processes or other highly expressed genes are likely to appear after sequencing about 1000 clones from a non-normalized cDNA library. Generally, high levels of expression indicate an important function in the organism. It may require more sequencing to obtain low expression genes or genes expressed only in a critical period, for example transcription factors. However, the highly expressed genes involved in energy metabolism, cellular defence and stress responses are important for homeostasis, and are thus potential candidates for sublethal markers against environmental stress and xenobiotics.

Applications

EST data contributes to the understanding of functional genes and gene networks, and has also been used for identification of non-protein coding mRNA with putative functions (Hirsch et al. 2006). Publicly available ESTs have also been used for subsequent novel phylogenetic analyses for species and groups (see subsection 'Genome sequencing: comparing genomes—phylogenomics'). Analysis of ESTs can also reveal the presence of microsatellite-containing genes, single nucleotide polymorphisms (SNPs), and other populational markers (Chen et al. 2006). Overall, the cDNA clones and EST sequence information (www.ncbi.nlm.nih.gov) are very useful for post-genomic functional analyses of the biology of the organism and for investigating links between evolution, ecology, physiology, genes and proteins.

Sequence information from an EST database can subsequently be used to quantify mRNA expression in a more focused experiment by gene-specific RT-PCR or other methods such as Northern blotting. *In situ* hybridization and antibody labelling can also reveal where in the organism the particular genes and proteins are expressed in both tissue sections and in whole animals. Identified genes coding for enzymes can, for example, be tested as putative novel biomarkers useful for simple enzyme activity based assays at the protein level.

The sequences can also be used for subsequent high throughput micro- or macroarray approaches (see subsection 'Microarrays/proteomics'), where clones or synthesized DNA oligos are arrayed for high throughput hybridization. Highly expressed housekeeping genes or structural genes that are likely to be obtained in an EST collection, such as actin ou 18s, can be used as controls in expression experiments. A good example of a study combining ESTs with subsequent expression studies at a smaller scale is that carried out by Gueguen et al. (2003), who first sequenced 1142 cDNA clones made from an enriched library of hemocytes from bacteria-challenged oysters. After annotating their sequences, they identified 20 genes with putative immune function. Subsequent expression studies of 4 of these genes then revealed that 3 of them were indeed induced by bacteria. In an ecotoxicological study, Nakayama et al. (2006) constructed a DNA oligo array of 1061 sequences using sequence information from an EST collection from the common cormorant. They hybridized this array with cDNA from livers obtained from wild cormorants and could correlate levels of certain environmental contaminants found in the animals with altered expression of P450 and antioxidant enzymes in the liver.

The blue mussel has been widely investigated in biomonitoring programs and recognized as a potential candidate species for marine genomic approaches in ecotoxicology (Wilson et al. 2005). An early EST project using multiple tissues from unstressed blue mussels revealed an expression profile and sequence data of known and unknown genes (Venier et al. 2003), and this information was recently used to design a low-density DNA oligo array for stress response detection (Dondero et al. 2006). Such small arrays may advance the use of genomics in marine biomonitoring.

In conclusion, ESTs provide sequence information useful for phylogenetic studies, population genetics, ecotoxicology, array projects and downstream gene- or protein-specific studies, all very useful for the understanding of the organism in its relationship with the environment. They can also be the starting point for more ambitious genome projects.

MICROARRAYS/PROTEOMICS

Methods

Microarrays

Transcription profiling using microarrays is expected to be the major activity of ecological genomics in the near future. This method allows analyses of the kinds and amounts of mRNA produced by a cell or tissue,

and therefore the facility to understand which genes are expressed. This in turn provides insights into how the cell/tissue responds when it grows or multiplies, changes function, or when it is subject to new or unnatural environmental conditions. Gene expression is a highly complex and tightly regulated process that allows a cell to respond dynamically both to environmental stimuli and to its own changing needs. This mechanism acts as both an 'on/off' switch to control which genes are expressed in a cell, and as a 'volume control' that increases or decreases the level of expression of particular genes as necessary. DNA microarray technology, in correlation with genome projects as well as phylogenetic and comparative genomic approaches, may also facilitate the identification and classification of DNA sequence information and the assignment of functions to newly identified genes (Wilson et al. 2006).

Common to all microarray approaches is the basic principle of complementary base pairing. A microarray operates by exploiting the ability of a given mRNA molecule to bind specifically and non-covalently to, or hybridize to, the DNA template from which it originated. By using a microarray, chip or slide, which consists of respective gene sequences or ESTs that are coated on a solid layer at high density, it is possible to determine, in a single experiment, the expression levels of hundreds or thousands of genes by measuring the amount of mRNA bound to each site on the array. The subsequent use of a computer driven microarray reader enables precise measurement of the amount of mRNA hybridized to the spots on the microarray. This generates a profile of gene expression for a cell or a cell population/tissue that can be used to build a molecular fingerprint. A judgement on the respective genes with regard to expression level is possible for distinct time points or response states. Moreover, besides qualitative assessment, the data also can be evaluated quantitatively, which may be highly relevant to both the ecological or ecotoxicological response of a species and its environmental management. Gene expression profiles thus provide a molecular fingerprint of the transcriptome. To date, ecologists have not used the global-gene expression response pattern per se as a 'signature response pattern' to changing conditions. Nevertheless, transcriptome pattern signatures, as a response to changing physiology, are increasingly used in medicine (in particular for diagnostic purposes), and it is only a matter of time before the approach crosses over to ecology (Chen et al. 2005, Jones et al. 2005, Selman et al. 2006).

To fabricate expression microarrays, EST complementary DNA (cDNA), or gene-specific sequences that are synthesized *in situ*, are spotted at defined positions on a surface (e.g. glass slide, nylon membrane). The mRNAs of interest (samples) and a control mRNA (reference) are then transformed into cDNAs, and each sample and the reference are labelled by different fluorochromes and co-hybridized (Fig. 3). The detection of the hybridization signals requires a specific microarray scanner connected to a database, which is essen-

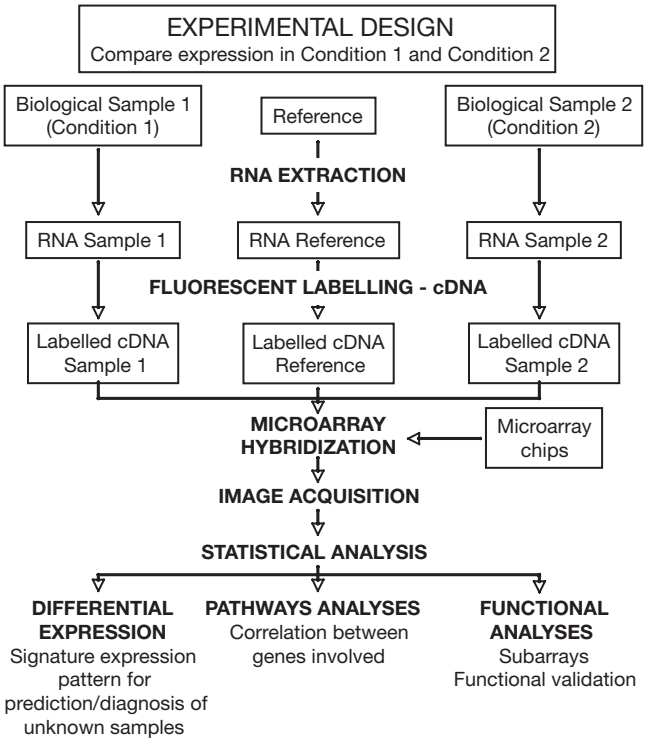

Fig. 3. Microarray experiment flowchart. cDNA microarrays are used in gene expression analysis. In this technique, RNA is isolated from 2 (or more) different samples derived from study subjects under different conditions, as well as from a common reference (which serves as a calibrator) (RNA EXTRACTION). The samples and common reference RNAs are transformed into cDNA and labelled by 2 fluorochromes respectively: 1 fluorochrome for the reference and 1 fluorochrome for the samples (generally the green cyanine 3 and the red cyanine 5 [Cy3, Cy5]) (FLUORESCENT LABELLING-cDNA). Each sample/common reference is then co-hybridised to a microarray replicate, which consists of large numbers of cDNAs/oligonucleotides orderly arranged/spotted onto a glass microscope slide (array spots) (MICROARRAY HYBRIDIZATION). After co-hybridization of each sample and common reference under stringent conditions on the microarray, a scanner records the emission of the 2 fluorochromes (for each sample-reference pair) for each spot on the microarray after excitation at given wavelengths (IMAGE ACQUISITION). The intensity of the fluorescence emission signals on each spot is proportional to transcript levels in the biological samples. For calibration purposes, the ratio of sample to reference emission for each microarray spot is used to compare the 2 (or more) study samples. Microarray data are then analysed using specific software that enables clustering of genes with similar expression patterns, which can be used to establish a differential expression signature for the samples compared (DIFFERENTIAL EXPRESSION). Responses of transcripts in various samples can also be clustered according to their affiliation with a particular intracellular signaling pathway (PATHWAY ANALYSES) or according to their common function (FUNCTIONAL ANALYSES)

tial for the analysis of the large amount of data. In addition, various algorithms must be applied to optimize the evaluation of the data. The objective is then to distinguish between random and significant patterns of gene expression among samples. After quality control of the sample and the hybridization comes image processing and the first analytical step to produce a large number of quantified gene expression values. These values represent absolute fluorescence signal intensities as a direct result of hybridization events on the array surface. The data are then normalized to compare the appropriate measured gene expression levels. The expression levels can then be used for several purposes. For example, it is possible to classify genes based on their expression levels in the different responses to the environment (e.g. environmental changes induce a number of genes to increase or decrease their expression) or to classify them in a functional way (e.g. all genes involved in cell membrane transport). Alternatively, if genes that change their expression belong to a biological network or pathway, they may be classified as such (e.g. genes involved in aerobic/anaeorobic cell respiration that change their behavior during anoxic conditions).

Proteomics

As microarrays can be used to assess changes in the transcriptome, proteomics can be used to study the proteome, that is all the proteins that are synthesized by a particular cell at a particular time. The proteome is the protein complement of the genome and the study of proteomics is important because proteins are responsible for both the structure and the functions of all living things, whereas genes are simply the instructions for making proteins. Moreover, the proteome more accurately reflects the response because post-translational and post-transcriptional modifications as well as phosphorylations etc. can substantially change the nature of the expressed protein product. The set of proteins within a cell varies both from one differentiated cell type to another (e.g. in development) and over time, depending on the activities of the cell (e.g. division during algal cell blooms; repairing damage to DNA when pollutants occur; responding to a newly available nutrient or stress factor when the environment changes; responding to the arrival of a hormone during mating season etc.). In this way, the proteome is a genuine measure of the cell phenotype.

In proteomics, protein mixtures are extracted from cells or tissues that have been exposed to an environmental condition or that represent a temporal or spatial condition (sample). At the same time other 'normal' cells or tissues are used as controls (control). Each type (sample and control) of protein mixture is subsequently subject to 2D gel electrophoresis, which separates the proteins in one dimension by their electrical charge and in the second dimension by their size. The gel is then stained to visualize various protein spots, and spots of sample and control gels are compared to identify differentially expressed proteins. Interesting (i.e. differentially expressed) spots are punched out of the gel, and analyzed. The analysis generally starts with treatment by a protease to digest the protein into a mix of peptides that can be run through a mass spectrometer to separate the peptides into sharply defined peaks. The result is mined against a database of all known proteins (which have been digested with the same enzyme) to see if a match can be found. If no match is found for the digested protein, a mass spectrometer can be used first to randomly break the peptide into a mix of fragments containing 1, 2 etc. amino acids and then to measure the mass of each fragment. The resulting data can be searched against a database that matches the mass data with known pairs, triplets etc. of amino acids. Subsequently overlapping fragments are assembled to reveal the entire sequence of the peptide. This can be searched against a genetic database to find the gene that encodes this particular peptide. In turn, translation of the matching gene reveals the entire sequence of the protein.

Another method frequently used to deliver valuable results for proteomics research is 2D nano-liquid chromatography-mass spectrometry (LC/MS). For instance, it has been used successfully in elucidating the proteome of several organisms (Washburn et al. 2001, Florens et al. 2002, Nägele et al. 2004).

Applications

DNA microarrays and proteomics have great potential to reveal community dynamics at different levels from individual genes to communities. This will become essential in population genetics and the analysis of biodiversity. These techniques are already being applied to marine ecology.

Large and medium environmental effects

The comprehensive description of transcriptomic responses provides useful information for conservation efforts, because it provides additional tools for early diagnostics. For example, a number of proteomic and genomic studies are underway to develop early markers for toxic algal bloom prediction (Chan et al. 2004, Lidie et al. 2005). Biomarkers for pollution in mussels are also being unveiled by proteomics, such as 2D gel

electrophoresis for peroxisome proliferation (Mi et al. 2005) or protein chip technology (Knigge et al. 2004). For restoration and bioregeneration efforts, genomics can help decipher the metabolic pathways involved in greenhouse gas balance in the ocean, such as those employed by the coccolithophore *Emiliania huxleyi*, which mediates oceanic and atmospheric carbon cycling (Nguyen et al. 2005, Dyhrman et al. 2006), or those used by methane-consuming bacteria (Hallam et al. 2004).

Ecotoxicology

By enabling the analysis of chemical effects at the molecular, tissue, and whole organism level, emerging technologies in the areas of genomics, proteomics, and metabolomics are important for the development of streamlined, cost-effective, and comprehensive testing approaches for evaluating environmental hazards. The genomic tools for ecotoxicogenomics have been reviewed by Wilson et al. (2005), and also recently by Miracle & Ankley (2005) with a particular emphasis on fish testing. Increasingly, more studies are emerging in this field, such as that of the effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) exposure on zebrafish caudal fin regeneration (Andreasen et al. 2006). Proteomics have also been used to follow protein profile alterations after exposure to cadmium in the marine alga *Nannochloropsis oculata* (Kim et al. 2005).

Adaptation and colonization of new habitats

Symbiotic associations are fundamental to the survival of many organisms, their diversity and even colonization of previously inhospitable habitats. Cnidarian-dinoflagellate intracellular symbioses are common mutualisms in the marine environment. They form the trophic and structural foundation of coral reef ecosystems and have played a key role in the radiation and biodiversity of cnidarian species. Proteomic studies to look at the interaction between host and symbiont have already begun (Barneah et al. 2006), as have systematic analyses of EST and cDNA microarray studies (Kuo et al. 2004, Rodriguez-Lanetty et al. 2006). This should ultimately lead to the discovery and characterization of symbiosis gene markers, which will enable early diagnosis of coral bleaching, a phenomenon that can ultimately lead to coral reef ecosystem breakdown owing to the loss of dinoflagellate symbionts from cnidarian hosts. One marker has currently been developed for the sea anemone *Anthopleura elegantissima* (Mitchelmore et al. 2002), but larger screenings will probably identify more.

Molecular responses permitting tolerance to extreme environments are also important to our understanding of how organisms have diversified and adapted. Certain halophilic archaea, for example, can develop anaerobic capabilities when high salt concentrations, elevated temperatures, and high cell densities promoted by aerobic growth and flotation reduce the availability of molecular oxygen. An operon with proteins responsible for and/or induced during anaerobic respiration was found in *Halobacterium* sp. by using transcriptome analysis as a complement to other methods such as phenotype analysis (Muller & DasSarma 2005).

The recent completion of several algal genome sequences and EST collections has facilitated functional genomic approaches for algal model systems. Ecological questions such as acquisition of increased metabolic versatility can be answered using these techniques. For example, the thermo-acidophilic unicellular red alga *Galdieria sulphuraria* can adopt heterotrophic and mixotrophic growth modes on more than 50 different carbon sources, and tolerate hot acidic environments as well as high concentrations of toxic metal ions, suggesting potential applications in bioremediation. To unravel the exceptional metabolic pathways of this organism, Weber et al. (2004) used a comparison between the *G. sulphuraria* transcriptome and the obligate photoautotrophic red alga *Cyanidioschyzon merolae*, which has a similar genome size. This study suggested that genes involved in the uptake of reduced carbon compounds and related enzymes were crucial to the metabolic flexibility of *G. sulphuraria* (Barbier et al. 2005). Proteomic approaches for dissecting molecular mechanisms of salinity tolerance in algae and higher plants are also in progress (Liska et al. 2004).

Ecophysiology and behavioural ecology

Atlantic salmon *Salmo salar* are known for spectacular marine migrations before homing to spawn in natal rivers. However, many males do not migrate before reproducing. Rather, these so-called 'sneaker' males mature early and reproduce at much smaller sizes than their migratory conspecifics without ever leaving freshwater. Early sexual maturity in salmon is the result of developmental plasticity, because the same genotype can express both types of reproduction tactics depending on the environment. Aubin-Horth et al. (2005) investigated the nature and extent of the coordinated molecular changes that accompany such a fundamental transformation by comparing brain transcriptional profiles of wild, mature sneaker males to age-matched, immature males (future large anadro-

mous males) and immature females. Of the ca. 3000 genes surveyed, 15% were differentially expressed in the brains of the 2 male types, and consistent patterns of gene expression were found for individuals of the same reproductive tactic. Notably, gene expression patterns in immature males differed from both immature females and sneakers, indicating that delayed maturation and sea migration by immature males, the 'default' life cycle, may actually result from an active inhibition of development into a sneaker. In this context, it is notable that a salmonid microarray containing cDNAs representing 16006 genes has been developed and assayed for intraspecific variation hybridization studies (von Schalburg et al. 2005).

A number of organisms change their activities and physiology during the circadian cycle: they emit chemical substances into the environment or bioluminesce, therefore influencing the ecosystem that they are part of. The first hints of temporal control within chloroplast proteins of *Arabidopsis thaliana* were identified by proteome analysis, and the technology has now been applied to the green alga *Chlamydomonas reinhardtii* (reviewed by Wagner et al. 2005), and chronobiological proteome assays have been performed for the dinoflagellate *Lingulodinium polyedra* (Akimoto et al. 2004).

Evolutionary ecological studies

The applications of functional genomics to evolutionary ecological studies have been explored by Feder & Mitchell-Olds (2003) and in the marine field were reviewed by Wilson et al. (2005), with a special focus on the plastic nature of the genome as seen by whole-genome comparisons. These applications also included assessment of influences on morphology and speciation brought about by variations in Quantitative Trait Loci (QTLs) and/or changes in non-coding regulatory sequences that control the expression of genes in time and space.

Limitations

Transcription profiling and proteomics are not ends in themselves and, for example, although very powerful and extremely useful, microarrays are simply exploratory instruments. This is only touching the surface of what ecological genomics has to offer. To be useful and worthwhile it needs to be grounded in physiological and biochemical knowledge, not to say understanding. It is rather more complicated than genomic analyses because, as noted earlier, a single gene can give rise to a number of different proteins through alternative splicing of the pre-messenger RNAs, RNA editing of the pre-messenger RNAs, and/or post-translational processing such as attachment of carbohydrate residues to form glycoproteins and addition of phosphate groups to some of the amino acids in the protein (Black 2000, Schmucker et al. 2000). There is a disparity between mRNA and protein abundance and enzyme activity, supporting the contention that it is difficult to predict protein activity from genomic data such as microarrays or RT-PCR (Glanemann et al. 2003). Moreover, some evidence suggests that there is no direct correlation between mRNA and protein changes with phenotype and fitness (Jeong et al. 2001, Giaever et al. 2002, Carpenter & Sabatini 2004). These observations are not surprising and can be explained by variability in mRNA stability, translational control, post-translational modifications and regulation of enzyme activity. Moreover, genes physically adjacent in the genome often have similar expression profiles when comparing different environments. Genes present in these expression clusters proved to be no more similar in structure or function than could be expected by chance, and are not expressed because they play a particular role but because a neighbour is expressed (Spellman & Rubin 2002). In this regard, genomics, transcriptomics and proteomics go hand in hand and perhaps ideally should be used in parallel to study the same processes.

These techniques will certainly play a key role in ecology, but only in combination with other emerging tools used to try to unravel the complex questions surrounding the question of how genomes interact with their environment. A fully detailed picture of the state of any biological system requires knowledge of all its components (i.e. transcriptome, proteome, and metabolome).

BARCODING

Method

One of the beneficial side effects of the genomic revolution is that not only has it helped the discovery of sequences of interest for population genetics (microsatellites, SNPs, etc.), but also the identification of species using DNA barcoding. The concept of DNA barcoding has attracted much attention from a wide range of biological disciplines (Lipscomb et al. 2003, Seberg et al. 2003, Stoeckle 2003, Janzen 2004, Marshall 2005) and offers intriguing perspectives for applications in marine ecology (Schander & Willassen 2005). The method allows systematic screening of one or several reference genes for as many organisms as is feasible (Hebert et al. 2003). If assembled into a comprehensive database, these sequences can then be used as refer-

ence genes for the identification of species based on sequence comparisons. Large-scale DNA barcoding libraries are already under construction, for example the 'Barcode of Life Data Systems' (BOLD, www.barcodinglife.org) and the 'Consortium for the Barcode Of Life' (CBOL, http://barcoding.si.edu), and proponents of the method envisage that in the future the ability to determine species will no longer depend on the taxonomic expertise of a few specialists. Instead, by simply obtaining a DNA sequence from the organism in question, anybody should be able to determine species identification. The barcoding idea is partly built upon the already common practice of including molecular data in taxonomic studies. Electrophoresis (Thorpe & Ryland 1979) or sequencing of nuclear genes (Floyd et al. 2002) has earlier been used to discriminate between morphologically indistinguishable/identical species. Likewise, in ecological surveys, genetic or proteomic markers have become essential for species determination, as for example in commercially important marine species (López et al. 2002) or toxic strains of algae (Chan et al. 2004, Lidie et al. 2005). The novel idea with barcoding sensu-stricto (Hebert et al. 2003), however, is to find a single marker that is universally applicable to a large group of organisms such as animals or plants, and for which general primers can be used. One of the proposed barcoding genes for metazoans is the mitochondrial gene cytochrome *c* oxidase subunit 1, also referred to as 'COI' or *'cox1'* (Fig. 4).

Applications

The ecological applications of a universal molecular identification system resulting from a marine barcoding program are vast and would improve the quality of ecological surveys tremendously, in particular in those that contain species difficult to identify. Such species are found everywhere in the marine realm, especially when entering micro- or meiofaunal assemblages. The majority of organisms on earth are microscopic with body sizes <1 mm and, although these play a central role in marine ecosystem function (Blaxter et al. 2005), most them (e.g. nematodes) are as yet undescribed. Also, amongst larger animals, in particular those with few diagnostic features, identification is a complex exercise usually restricted to experts, e.g. platyhelminths, nemerteans, or nematodes. Schander & Willassen (2005) showed that major parts of the faunal composition in marine reports and inventories often remain undetermined. This greatly impairs the comprehensibility of such studies and limits the conclusions that can be drawn.

There are many more potential applications arising from a marine barcoding program. Principally, it should be possible to determine species from all kinds of life-history stages, for example eggs or planktonic larvae. Stomach contents could also be utilised in order to resolve food webs in marine ecosytems. Also, faunal remains on the sea floor may be traced back to their living origins. Moreover, parasitic or other symbiotic relationships can be described (Tops & Okamura 2003) without the need to identify the symbiont visually. It seems there is great potential in characterizing faunal assemblages in such a detailed fashion. Further applications of the method lie in conservation and management efforts, for example in the monitoring of invasive species (DeSalle & Amato 2004). Dispersal in the marine environment is less hampered by geographical barriers compared with most terrestrial or limnic systems (Palumbi 1992), and invasive species are becoming an increasingly problematic side effect of globalization (Roman & Palumbi 2004). Here, barcodes could assist the tracing of invasive species, for example by scanning water samples and screening for the species in question.

Because of high variation at the species level, barcoding genes can also be used for other applications and vice versa. COI for example is a frequently applied marker in phylogeographic and phylogenetic studies. Hence, the genes used for species identification also have the potential to show the presence of cryptic species (Obst et al. 2005) and polymorphisms (Eriksson et al. 2006), describe the population structure within a species (Barber et al. 2002, Lessios et al. 2003), and test hypotheses of evolutionary relationships (Sorensen et

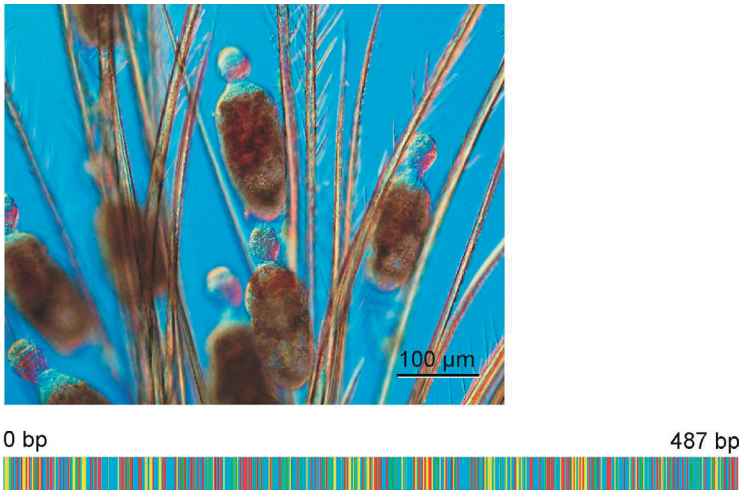

Fig. 4. A COI DNA-barcode (below) of the microscopic Cycliophora (above), an enigmatic protostome, so far only described from the mouth appendages of 3 nephropid lobsters (from Obst et al. 2005)

al. 2006). Thus, there is great potential for application of a marine barcoding program, and there are many examples of how genomic methods stimulate traditional fields such as taxonomy and ecology as well as training for a new generation of marine biologists with expertise in integrative approaches (Will et al. 2005).

Limitations

Recently, a number of examples have shown the ability of this technique to assign previously unidentified individuals to the right species using COI barcodes of (Hebert et al. 2003, 2004a, Hogg & Hebert 2004, Barrett & Hebert 2005) as well as the potential to discover new species (Hebert et al. 2004b, Obst et al. 2005). However, some principal issues regarding the method remain. For example, a recent study by Meyer & Paulay (2005) showed that the technique is successful only in those cases where the studied taxa are well known beforehand. This study revealed high error rates in the determination of species when the group was little studied. This is certainly the case for most of the biological diversity in the marine realm (May 1988). Furthermore, barcodes not only vary among but also within each species. This means that in order to reliably assign a sequence to the correct species, any reference library must take into account the entire intra-specific variation of the marker. Practically, this dictates the assortment of many barcodes necessary for each species, e.g. more than 10 sequences from the entire distribution area (Hajibabaei et al. 2005). Such problems will lead barcoding programs beyond the assembly of sequence information for a large number of organisms. Automatically, any barcoding project will necessitate parallel revisions of the groups under study, e.g. all sequence records need to be linked to voucher specimens that represent a species already described in considerable detail with regard to its taxonomy, morphology, and ecology (Hajibabaei et al. 2005).

ECOLOGY FEEDBACK TO GENOMICS

The genomics revolution provides some striking new insights for ecological studies. However, genomics is more than a toolbox added to marine ecology. Marine ecological genomics is new discipline merging genomics with marine ecology and leads to new questions independent of both fields. Despite such potential, genomic techniques present some limitations (e.g. see 'Genome sequencing: comparing genomes — phylogenomics' and 'Barcoding: limitations') that highlight the parallel importance of traditional taxonomy and ecological approaches.

One interesting example comes from the worm-like marine animal *Xenoturbella* spp. Even though they are neither parasitic nor microscopic, they lack a through-gut, gonads, coelomic cavities, and a brain. Owing to a simple body plan, their phylogenetic position has long remained puzzling. Based on morphology, they have been suggested to be a primitive flatworm (Westblad 1949), unique representatives of a plesiomorphic metazoan group (Jagersten 1959), an enteropneust, holothurian, or unique representatives of a deuterostome group (Reisinger 1960), a hemichordate (Pedersen & Pedersen 1986), an acoel flatworm (Franzen & Afzelius 1987, Lundin 1998, 2000, 2001), a primitive metazoan (Ehlers & Sopott-Ehlers 1997, Raikova et al. 2000), a bivalve (Israelsson 1997, 1999), or a bryozoan (Zrzavy 1998). This example shows that, even when using rather powerful techniques such as scanning or transmission electron microscopy, morphology alone cannot resolve the phylogenetic position of an animal.

In 1997, a gene sequence analysis of the mitochondrial *cox1* gene showed that *Xenoturbella* spp. are bivalves (Norèn & Jondelius 1997), solving the problem at last. However, in 2003, Bourlat et al. reported a different sequence of the same *cox1* gene from *Xenoturbella* spp., and suggested that they were deuterostomes. Are *Xenoturbella* spp. bivalves, or are they deuterostomes? This is a good example of where gene sequence analysis, despite wide acceptance today, is not all-powerful, and cannot alone determine the phylogenetic position of a simple animal.

Bourlat et al. (2003) showed that if you extract DNA from the epidermis alone, you obtain mostly deuterostome sequence. This suggests that the molluscan DNA was in the gut and was that of the prey of *Xenoturbella* spp.. This hypothesis clearly needs to be tested, and it is here that ecological data becomes useful, if not essential. The reported bivalve sequence showed 97.2% homology to *Nucula tenuis* at the nucleotide level. Through ecological projects such as the national Swedish monitoring program (Agrenius 2003), it is known that *N. tenuis* are present in the fjord where *Xenoturbella* spp. are found. Furthermore, the reported deuterostome sequence shows no match with other deuterostome animals in that area. These ecological and molecular data together support the contention that the bivalve DNA is that of *N. tenuis* in the area where *Xenoturbella* spp. feed, and that the deuterostome sequence is the genuine *Xenoturbella* spp. sequence. An immunohistochemical study has also supported the deuterostome status of *Xenoturbella* spp. (Stach at al. 2005). Thus, because the identity question has now gained morphological, gene sequence, and immunohistochemical evidential support together with ecological data, the answer to the long-lasting question has

finally been resolved: *Xenoturbella* spp. are deuterostomes.

Even when not investigating such a complicated example as *Xenoturbella* spp., it is vital to avoid contamination when applying genomic methods. This includes contamination from prey, parasites, animals attached to your intended animal, or contamination during experimental procedures in the lab. The first 3 factors are especially important when working with non-model organisms collected from nature, and highlights the importance of understanding ecology. Furthermore, there are always possibilities of artefacts, such as the formation of a concatemer of 2 separate genes during gene cloning (Hibino et al. 2004) or incorrect PCR priming (Quist & Chapela 2001, Metz & Futterer 2002, Kaplinsky et al. 2002), so thorough analyses, repeat sequencing and cross-checking are a necessity in genomic studies.

DISCUSSION

Previous ecological molecular studies have focused on limited numbers of genes and gene products. To understand complex life processes, a more integrative approach is necessary and 'approaches similar in spirit to systems biology should ultimately be adopted to enable genomics answers to ecological questions' (Van Straalen & Roelofs 2006). Recent advances in molecular techniques have made high throughput analyses of genomes, transcriptomes and proteomes possible and with this, ecology has entered a new era. Nevertheless, the incredibly powerful engine called genomics is still in its infancy and its inductive phase. One common criticism of current massive data-collection efforts is that much information, but little knowledge, is accumulating. This descriptive and not hypothesis-driven phase can be a source of impressive data sets and often unexpected information, and new hypotheses may be derived. The next step will be a more integrative approach and hypothesis-driven science. It will allow us to answer deep biological and evolutionary questions linking spatial and temporal considerations with the interaction between genome and the environment. As suggested by James Galagan, 'It's no longer enough to sequence a genome, catalogue the genes and come up with diagrams of signaling and so forth. We're expecting to get much more'.

What kind of evolution can we expect in the near future for marine ecological genomics? (1) At present, there are few sequenced genomes of ecologically relevant species in the marine environment. Technological advances in the near future will allow an increase in the number of these species and allow genome-wide analyses of ecological questions. (2) Metagenomics approaches are particularly promising in ecology, and we can expect reconstruction of complete genomes from large-scale sequencing of the environment. (3) One limitation is that the great majority of genomics studies are conducted in the laboratory (perhaps with the exception of microbial ecology), so analysis performed directly in the field will allow us to answer new questions. (4) We can also expect some development of new methods for data analysis. (5) The study of epigenetic variants in natural populations has little influence in ecology now, but it will eventually have more impact thanks to 'omics' technologies (Van Straalen & Roelofs 2006).

This evolution is also a human challenge. Sequencing and analyzing a genome requires almost as many management skills as scientific ones. It often involves a large number of groups and therefore needs careful coordination (e.g. organization of conferences, workshops) between teams with different skills and goals. Moreover, it is an informatic challenge, and communication is crucial for the establishment of standards, tools and algorithms, for example for the annotation of environmental genomic data. The real challenge for marine ecological genomics is the creation of sufficiently large but effective collaborative networks around key model species. Some networks are already devoted to the development, utilization, and spreading of 'omics' approaches for the investigation of the biology and ecology of marine organisms. A marine genomics project is a functional genomics initiative developed in the USA to provide a pipeline for the curation of ESTs and gene expression microarray data for marine organisms (46000 ESTs from 19 species in the database; see www.marinegenomics.org). It has provided a clearing-house for marine-specific EST and microarray data available online (McKillen et al. 2005). In Europe, the Network of Excellence 'Marine Genomics Europe' (MGE; see www.marine-genomics-europe.org) is a major new enterprise funded by the European Community, comprising 44 laboratories and standing the crossroads between life sciences, ecology, environment, bioinformatics and high technologies within a multicultural European environment. MGE is devoted to the development, utilization, and spreading of high-throughput approaches for the investigation of the biology marine organisms. Uniquely, it has enabled the integration of a hitherto fragmented set of high-level expert groups to come together, share skills, state-of-the-art platforms and ambition. Benefits include large-scale sequencing projects, phylogenetic analyses, and the application of genomics technologies to functional, comparative, and environmental issues in marine biology. Thus, while marine ecological genomics is not completely beyond the reach of individuals, there are clear advantages to be gained from

formal, or informal, consortia brought together to solve common issues and shared ideals. The increasing need for multidisciplinarity, combined with the costs of capital equipment and associated resources, makes such networks an important component for the future development of marine genomics, not least by providing opportunities for training the next generation of scientists and enabling the creation of sustainable collaborations.

Acknowledgements. This work is supported by VR (Swedish Research Council) EU RTNZ-2001-00029, Network of Excellence Marine Genomics Europe GOCE-04-505403 and the Royal Swedish Academy of Sciences.

LITERATURE CITED

Adoutte A, Balavoine G, Lartillot N, de Rosa R (1999) Animal evolution—the end of the intermediate taxa? Trends Genet 5:104–108

Agrenius S (2003) Overvakning av mjukbottenfaunan langs Sveriges vastkust: rapport fran verksamheten ar 2002. Naturvardsverket 1–7, available at: http://www.marecol.gu.se/digitalAssets/751833_Rapport_2003.pdf

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropodes and other moulting animals. Nature 387:489–493

Akimoto H, Wu C, Kinumi T, Ohmiya Y (2004) Biological rhythmicity in expressed proteins of the marine dinoflagellate *Lingulodinium polyedrum* demonstrated by chronological proteomics. Biochem Biophys Res Commun 315:306–12

Andreasen EA, Mathew LK, Tanguay RL (2006) Regenerative growth is impacted by TCDD: gene expression analysis reveals extracellular matrix modulation. Toxicol Sci 92: 254–269

Aubin-Horth N, Landry CR, Letcher BH, Hofmann HA (2005) Alternative life histories shape brain gene expression profiles in males of the same population. Proc Biol Sci 272: 1655–62

Barber PH, Palumbi SR, Erdmann MV, Moosa MK (2002) Sharp genetic breaks among populations of *Haptosquilla pulchella* (Stomatopoda) indicate limits to larval transport: patterns, causes, and consequences. Mol Ecol 11:659–674

Barbier G, Oesterhelt C, Larson MD, Halgren RG, Wilkerson C, Garavito RM, Benning C, Weber AP (2005) Comparative genomics of two closely related unicellular thermoacidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae. Plant Physiol 137:460–474

Barneah O, Benayahu Y, Weis VM (2006) Comparative proteomics of symbiotic and aposymbiotic juvenile soft corals. Mar Biotechnol 8:11–16

Barret RDH, Hebert PDN (2005) Identifying arachnids through DNA sequences. Can J Zool 83:481–491

Béjà O (2004) To BAC or not to BAC: marine ecogenomics. Curr Opin Biotech 15:187–190

Béjà O, Aravind L, Koonin EV, Suzuki MT and 7 others (2000) Bacterial rhodopsin: evidence for a new type of phototropy in the sea. Science 289:1902–1906

Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 103:367–370

Blair JE, Ikeo K, Gojobori T, Hedges SB (2002) The evolutionary position of nematodes. BMC Evol Biol 2:7

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Phil Trans R Soc Lond B 360: 1935–1943

Bourlat SJ, Nielsen C, Lockyer AE, Littlewood DTJ, Telford MJ (2003) *Xenoturbella* is a deuterostome that eats molluscs. Nature 424:925–928

Carpenter AE, Sabatini DM (2004) Systematic genome-wide screens of gene function. Nature Rev Genet 5:11–22

Chan LL, Hodgkiss IJ, Wan JM, Lum JH, Mak AS, Sit WH, Lo SC (2004) Proteomic study of a model causative agent of harmful algal blooms, *Prorocentrum triestinum* II: the use of differentially expressed protein profiles under different growth phases and growth conditions for bloom prediction. Proteomics 4:3214–3226

Chen ZM, Crone KG, Watson MA, Pfeifer JD, Wang HL (2005) Identification of a unique gene expression signature that differentiates hepatocellular adenoma from well-differentiated hepatocellular carcinoma. Am J Surg Pathol 29:1600–1608

Chen C, Zhou P, Choi YA, Huang S, Gmitter FG (2006) Mining and characterizing microsatellites from citrus ESTs. Theor Appl Genet 11:1–10

Curtis TP, Sloan WT, Scannell JW (2002) Estimatic prokaryotic diversity and its limits. Proc Natl Acad Sci USA 99: 10494–10499

DeLong EF, Karl DM (2005) Genomic perspectives in microbial oceanography. Nature 437:336–342

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nature Rev Genet 6: 361–375

DeSalle R, Amato G (2004) The expansion of conservation genetics. Nature Rev Genet 5:702–712

Dondero F, Piacentini L, Marsano F, Rebelo M, Vergani L, Venier P, Viarengo A (2006) Gene transcription profiling in pollutant exposed mussels (*Mytilus* spp.) using a new low-density oligogonucleotide microarray. Gene 376: 24–36

Doney SC, Abbott MR, Cullen JJ, Karl DM, Rothstein L (2004) From genes to ecosystems: the ocean's new frontier. Front Ecol Environ 2:457–466

Dyhrman ST, Haley ST, Birkeland SR, Wurch LL, Cipriano MJ, McArthur AG (2006) Long serial analysis of gene expression for gene discovery and transcriptome profiling in the widespread marine coccolithophore *Emiliania huxleyi*. Appl Environ 72:252–60

Ehlers U, Sopott-Ehlers B (1997) Ultrastructure of the subepidermal musculature of *Xenoturbella bocki* , the adelphotaxon of the Bilateria. Zoomorphology 117:71–79

Eriksson R, Nygren A, Sundberg P (2006) Genetic evidence of phenotypic polymorphism in the aeolid nudibranch *Flabellina verrucosa* (M. Sars, 1829) (Opisthobranchia: Nudibranchia). Org Div Evol 6:71–76

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nature Rev Genet 4:649–655

Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool 27: 401–410

Floyd R, Abebe E, Papert A, Blaxter M (2002) Molecular barcodes for soil nematode identification. Mol Ecol 11: 839–850

Franzen A, Afzelius BA (1987) The ciliated epidermis of

Xenoturbella bocki (Platyhelminthes Xenoturbellida) with some phylogenetic considerations. Zool Scr 16:9–17

García-Fernàndez JM, Tandeau de Marsac N, Diez J (2004) Streamlined regulation and gene loss as adaptative mecanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. Microbiol Mol Biol Rev 68:630–638

Giaever G, Chu AM, Ni L, Connelly C and 69 others (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418:387–391

Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. Nature 437:343–348

Glanemann C, Loos A, Gorret N, Willis LB, O'Brien XM, Lessard PA, Sinskey AJ (2003) Disparity between changes in mRNA abundance and enzyme activity in *Corynebacterium glutamicum*: implications for DNA microarray analysis. Appl Microbiol Biotechnol 61:61–68

Gueguen Y, Cadoret JP, Flament D, Barreau-Roumiguiere C and 5 others (2003) Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas*. Gene 303:139–145

Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Bhattacharya D (2005) Insights into a dinoflagellate genome through expressed sequence tag analysis. BMC Genomics 6:80

Hajibabaei M, DeWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM, Hebert PDN (2005) Critical factors for assembling a high volume of DNA barcodes. Phil Trans R Soc Lond B 360:1959–1967

Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. Science 305:1457–1462

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond B 270:313–321

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004a) Identification of birds through DNA barcodes. PLoS Biol 2: 1657–1663

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004b) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc Natl Acad Sci USA 101:14812–14817

Held M, Gase K, Baldwin IT (2004) Microarrays in ecological research: a case study of a cDNA microarray for plant-herbivore interactions. BMC Ecol 4:13

Hess WR (2004) Genome analysis of marine photosynthetic microbes and their global role. Curr Opin Biotech 15:191–198

Hibino T, Harada Y, Minokawa T, Nonaka S, Amemiya S (2004) Molecular heterotopy in the expression of *Brachyury orthologs* in order Clypeasteroida (irregular sea urchins) and order Echinoida (regular sea urchins). Dev Genes Evol 214:546–558

Hirsch J, Lefort V, Vankersschaver M, Boualem A, Lucas A, Thermes C, d'Aubenton-Carafa Y, Crespi M (2006) Characterization of 43 non-protein coding mRNA genes in *Arabidopsis* including the MIR162a-derived transcripts. Plant Physiol 140:1192–1204

Hogg ID, Hebert PDN (2004) Biological identification of springtails (Collembola: Hexapoda) from the Canadian Arctic, using mitochondrial DNA barcodes. Can J Zool 82: 749–754

Israelsson O (1997) ...and molluscan embryogenesis. Nature 390:32

Israelsson O (1999) New light on the enigmatic *Xenoturbella* (phylum uncertain): ontogeny and phylogeny. Proc R Soc Lond B 266:835–841

Jagersten G (1959) Further remarks on the early phylogeny of Metazoa. Zool Bidr Upps 33:79–108

Janzen DH (2004) Now is the time. Phil Trans R Soc Lond B 359:731–732

Jeong H, Mason SP, Barabàsi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42

Jones M, Blaxter M (2005) Evolutionary biology: animal roots and shoots. Nature 434:1076–1077

Jones J, Otu H, Spentzos D, Kolia S and 8 others (2005) Gene signatures of progression and metastasis in renal cell cancer. Clin Cancer Res 11:5730–5739

Kaplinsky N, Braun D, Lisch D, Hay A, Hake S, Freeling M (2002) Maize transgene results in Mexico are artefacts. Nature 416:601

Kim YK, Yoo WI, Lee SH, Lee MY (2005) Proteomic analysis of cadmium-induced protein profile alterations from marine alga *Nannochloropsis oculata*. Ecotoxicology 14:589–596

Knigge T, Monsinjon T, Andersen OK (2004) Surface-enhanced laser desorption/ionization-time of flight-mass spectrometry approach to biomarker discovery in blue mussels (*Mytilus edulis*) exposed to polyaromatic hydrocarbons and heavy metals under field conditions. Proteomics 4:2722–2727

Kore-eda S, Cushman MA, Akselrod I, Bufford D, Fredrickson M, Clark E, Cushman JC (2004) Transcript profiling of salinity stress responses by large-scale expressed sequence tag analysis in *Mesembryanthemum crystallinum*. Gene 341:83–92

Kuo J, Chen MC, Lin CH, Fang LS (2004) Comparative gene expression in the symbiotic and aposymbiotic *Aiptasia pulchella* by expressed sequence tag analysis. Biochem Biophys Res Commun 318:176–186

Lawton JH (1994) What do species do in ecosystems? Oikos 71:367–374

Lessios HA, Kane J, Robertson DR (2003) Phylogeography of the pantropical sea urchin Tripneustes: contrasting patterns of population structure between oceans. Evol Int J Org Evol 57:2026–2036

Lidie KB, Ryan JC, Barbier M, Van Dolah FM (2005) Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of DNA microarray. Mar Biotechnol 7: 481–493

Lipscomb D, Platnick N, Wheeler Q (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol Evol 1865–66

Liska AJ, Shevchenko A, Pick U, Katz A (2004) Enhanced photosynthesis and redox energy production contribute to salinity tolerance in Dunaliella as revealed by homology-based proteomics. Plant Physiol 136:2806–2817

López JL, MarinaA, Alvarez G, Vazquez J (2002) Application of proteomics for fast identification of species-specific peptides from marine species. Proteomics 2: 1658–1665

Lundin K (1998) The epidermal ciliary rootlets of *Xenoturbella bocki* (Xenoturbellida) revisited: new support for a possible kinship with the Acoelomorpha (Platyhelminthes). Zool Scr 27:263–270

Lundin K (2000) Phylogeny of the Nemertodermatida (Acoelomorpha, Platyhelminthes). A cladistic analysis. Zool Scr 29:17–27

Lundin K (2001) Degenerating epidermal cells in *Xenoturbella bocki* (phylum uncertain), Nemertodermatida and Acoela (Platyhelminthes) Belg J Zool 131:153–157

Margulies M, Egholm M, Altman WE, Attiya S and 52 others (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Marshall E (2005) Taxonomy—will DNA bar codes breathe life into classification? Science 307:1037
May RM (1988) How many species are there on earth? Science 241:1441–1449
McKillen DJ, Chen YA, Chen C, Jenny MJ and 7 others (2005) Marine genomics: a clearing-house for genomics and transcriptomic data of marine organisms. BMC Genomics 6:34
McKusick VA, Ruddle FH (1987) A new discipline, a new name, a new journal. Genomics 1:1–2
Metz M, Fuetterer J (2002) Suspect evidence of transgenic contamination. Nature 416:600–601
Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS Biol 3:2229–2238
Mi J, Orbea A, Syme N, Ahmed M, Cajaraville MP, Cristobal S (2005) Peroxisomal proteomics, a new tool for risk assessment of peroxisome proliferating pollutants in the marine environment. Proteomics 5:3954–3965
Miracle AL, Ankley GT (2005) Ecotoxicogenomics: linkages between exposure and effects in assessing risks of aquatic contaminants to fish. Reprod Toxicol 19:321–326
Mitchelmore CL, Schwarz JA, Weis VM (2002) Development of symbiosis-specific genes as biomarkers for the early detection of cnidarian-algal symbiosis breakdown. Mar Environ Res 54:345–349
Muller JA, DasSarma S (2005) Genomic analysis of anaerobic respiration in the archaeon *Halobacterium* sp. strain NRC-1: dimethyl sulfoxide and trimethylamine N-oxide as terminal electron acceptors. J Bacteriol 187:1659–1667
Mushegian AR, Garey JR, Martin J, Liu LX (1998) Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res 8: 590–598
Nägele E, Vollmer M, Hörth P (2004) Improved 2D Nano-LC/MS for proteomics applications: a comparative analysis using yeast proteome. J Biomol Tech 15:134–143
Nakayama K, Iwata H, Kim EY, Tashiro K, Tanabe S (2006) Gene expression profiling in common cormorant liver with a oligo array: assessing the potential toxic effects of environmental contaminants. Environ Sci Technol 40: 1076–1083
Nguyen B, Bowers RM, Wahlund TM, Read BA (2005) Suppressive subtractive hybridization of and differences in gene expression content of calcifying and noncalcifying cultures of *Emiliania huxleyi* strain 1516. Appl Environ Microbiol 71:2564–2575
Norèn M, Jondelius U (1997) *Xenoturbella*'s molluscan relatives. Nature 390:31–32
Obst M, Funch P, Giribet G (2005) Hidden diversity and host specificity in cycliophorans: a phylogeographic analysis along the North Atlantic and Mediterranean Sea. Mol Ecol 14:4427–4440
Ogasawara M, Sasaki A, Metoci H, Shin-I T, Kohara Y, Satoh N, Satou Y (2002) Gene expression profiles in young adult *Ciona intestinalis*. Dev Gen Evol 212:173–185
Palumbi SR (1992) Marine speciation on a small planet. Trends Ecol Evol 7:114–118
Pedersen KJ, Pedersen LR (1986) Fine structural observations on the extracellular matrix (ECM) of *Xenoturbella bocki* Westblad 1949. Acta Zool 67:103–113
Peterson KJ, Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: interferences from morphology and 18S rDNA gene sequences. Evol Dev 3:170–205
Philippe H, Lartillot N, Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Edzysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246–1253
Pieper DH, Martins do Santos VAP, Golyshin PN (2004) Genomics and mechanistic insights into the biodegradation of organic pollutants. Curr Opin Biotech 15:215–224
Quist D, Chapela IH (2001) Transgenic DNA introgressed into traditional maize landraces in Oaxaca, Mexico. Nature 414:541–543
Raikova OI, Reuter M, Jondelius U, Gustafsson MKS (2000) An immunocytochemical and ultrastructural study of the nervous and muscular systems of *Xenoturbella westbladi* (Bilateria inc. sed.). Zoomorphology 120:107–118
Reisinger E (1960) Was ist *Xenoturbella*? Z Wiss Zool 164: 188–198
Rodriguez-Lanetty M, Phillips WS, Weis VM (2006) Transcriptome analysis of a cnidarian—dinoflagellate mutualism reveals complex modulation of host gene expression. BMC Genomics 7:23
Rogers YH, Venter JC (2005) Genomics: massively parallel sequencing. Nature 437:326–327
Roman J, Palumbi SR (2004) A global invader at home: population structure of the green crab, *Carcinus maenas*, in Europe. Mol Ecol 13:2891–2898
Roy SW, Gilbert W (2005) Complex early genes. Proc Natl Acad Sci USA 102:4403–4408
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467
Schander C, Willassen E (2005) What can biological barcoding do for marine biology? Mar Biol Res 1:79–83
Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell 101:671–684
Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends Ecol Evol 18:63–65
Selman M, Pardo A, Barrera L, Estrada A and 5 others (2006) Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis. Am J Respir Crit Care Med 173:188–198
Simon A, Glockner G, Felder M, Melkonian M, Becker B (2006) EST analysis of the scaly green flagellate *Mesostigma viride* (Streptophyta): Implications for the evolution of green plansts (Viridiplantae). BMC Plant Biol 6:2
Smith LM, Sanders, JZ, Kaiser RJ, Hugues P and 5 others (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679
Sneddon LU, Margareto J, Cossins AR (2005) The use of transcriptomics to address questions in behaviour: production of a suppression subtractive hybridisation library from dominance hierarchies of rainbow trout. Physiol Biochem Zool 78:695–705
Sorensen MV, Sterrer W, Giribet G (2006) Gnathostomulid phylogeny inferred from a combined approach of four molecular loci and morphology. Cladistics 22:32–58
Spellman PT, Rubin GM (2002) Evidence for large domains of similarity expressed genes in the *Drosophila* genome. J Biol 1:5
Stach T, Dupont S, Israelson O, Fauville G, Nakano H, Kanneby T, Thorndyke M (2005) Nerve cells of *Xenoturbella bocki* (phylum uncertain) and *Harrimania kupfferi* (Enteropneusta) are positively immunoreactive to antibodies raised against echinoderm neuropeptides. J Mar Biol Assoc UK 85:1519–1524
Steele HL, Streit WR (2005) Metagenomics: advances in ecology and biotechnology. FEMS Microbiol Lett 247:105–111

Stoeckle M (2003) Taxonomy, DNA, and the bar code of life. Bioscience 53:796–797

Strous M, Pelletier E, Mangenot S, Rattei T and 33 others (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440:790–794

Thorpe JP, Ryland JS (1979) Cryptic speciation detected by biochemical genetics in 3 ecologically important intertidal bryozoans. Estuar Coast Mar Sci 8:395–398

Tops S, Okamura B (2003) Infection of bryozoans by *Tetracapsuloides bryosalmonae* at sites endemic for salmonid proliferative kidney disease. Dis Aquat Org 57:221–226

Tringe SG, von Mering C, Kobayashi A, Salamov AA and 9 others (2005) Comparative metagenomics of microbial communities. Science 308:554–557

Van Straalen NM, Roelofs D (2006) An introduction to ecological genomics. Oxford University Press, Oxford

Venier P, Pallavicini A, De Nardi B, Lanfranchi G (2003) Towards a catalogue of genes transcribed in multiple tissues of *Mytilus galloprovincialis*. Gene 314:29–40

von Schalburg KR, Rise ML, Cooper GA, Brown GD, Gibbs AR, Nelson CC, Davidson WS, Koop BF (2005) Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. BMC Genomics 6:126

Wagner V, Gessner G, Mittag M (2005) Functional proteomics: a promising approach to find novel components of the circadian system. Chronobiol Int 22:403–415

Washburn MP, Wolters D, Yates III JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnol 19: 242–247

Watanabe H, Tatarazuko N, Oda S, Nishide H, Uchiyama I, Morita M, Iguchi T (2005) Analysis of expressed sequence tags of the water flea *Daphnia magna*. Genome 48: 606–609

Weber AP, Oesterhelt C, Gross W, Brautigam A and 13 others (2004) EST-analysis of the thermo-acidophilic red microalga *Galdieria sulphuraria* reveals potential for lipid A biosynthesis and unveils the pathway of carbon export from rhodoplasts. Plant Mol Biol 55:17–32

Westblad E (1949) *Xenoturbella bocki* n.g., n.sp., a peculiar, primitive turbellarian type. Arkiv Zool 1:3–29

White TJ (1996) The future of PCR technology: diversification of technologies and applications. Tibtech 14:478–483

Will KW, Mishler BD, Wheeler QD (2005) The perils of DNA barcoding and the need for integrative taxonomy. Syst Biol 54:844–851

Wilson K, Thorndyke M, Nilsen F, Rogers A, Martinez P (2005) Marine systems: moving into the genomics era. Mar Ecol 26:3–16

Wilson AC, Dunbar HE, Davis GK, Hunter WB, Stern DL, Moran NA (2006) A dual-genome microarray for the pea aphid, *Acyrthosiphon pisum*, and its obligate bacterial symbiont, *Buchnera aphidicola*. BMC Genomics 7:50

Winnepenninckx B, Backeljau T, Mackey LY, Brooks JM, de Wachter R, Kumar S, Garey JR (1995) 18S rRNA data indicate that Aschelminthes are polyphyletic in origin and consists at least three distinct clades. Mol Evol Biol 12: 1132–1137

Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res 14:29–36

Zrzavy J, Mihulka S, Kepka P, Bezdek A, Tietz D (1998) Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. Cladistics 14:249–285

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: July 5, 2006; Accepted: October 13, 2006
Proofs received from author(s): February 19, 2007

Vol. 332: 275–279, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Applications of proteomics in marine ecology

J. L. López*

Departamento de Genética, Facultad de Biología, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain

ABSTRACT: Proteomics emerged in the beginning of the 1990s due to the need for new methods for protein analysis. Proteomics is a much newer discipline than genomics, and confronts similar challenges to those that genomics researchers faced in the implementation of large-scale sequencing programs. The definition of proteomics as the use of quantitative protein-level measurements of gene expression to characterize biological processes and decipher the mechanisms of gene expression control fits in with any biological approach. In the present study, proteomics is discussed and defined in parallel with genomics, given that many authors integrate proteomics in the context of functional genomics. In this Theme Section, several facets of proteomics in marine ecology were adressed: capacity, or what can be done, utility, the technology possibilities, and how to use the data obtained. As with any new and interesting technology, the expectations often exceed reality. The applications of proteomics, the advantages and disadvantages, as well as a few limitations are discussed.

KEY WORDS: Proteomics · Marine ecology · Two-dimensional electrophoresis · Liquid chromatography · Mass spectrometry

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

The sequencing of complete genomes and the development of large expressed sequence tag (EST) databases have provided us with an understanding of the genomic capacity of many organisms. However, by themselves, these data are of limited use when it comes to fully understanding processes such as development, physiology and environmental adaptation. In order to understand these processes, scientists are now faced with the problem of how to best study the co-expression of large numbers of genes under biologically meaningful conditions. Large-scale gene expression studies can be conducted using either genomic (nucleic acid-based) or proteomic (protein-based) approaches. Genomics has moved into this functional phase through the advent of technologies such as microarrays and gene probes, used to detect gene activity through messenger display (Debouck & Goodfellow 1999). These technologies have been used to produce large-scale data sets that contain information about the messenger RNA (mRNA) molecules or 'transcripts' that are present in a cell, tissue or organism at a particular time (the transcriptome), and have led to a field of study referred to as transcriptomics.

The field of proteomics involves the study of proteomes. The term 'proteome' was originally defined as all the proteins expressed by the genome (Wilkins et al. 1996). However, it is now accepted that the proteome of an organism is more than simply a catalogue of all proteins encoded by the genome because it also includes the dynamic changes within the proteome, such as post-translational modifications that occur in response to various stimuli. An area of study within proteomics is 'expression proteomics', which is defined as the use of quantitative protein-level measurements of gene expression to characterize biological processes and decipher the mechanisms of gene expression control (Anderson & Anderson 1982). Expression proteomics allows researchers to obtain a quantitative description of protein expression and its changes under the influence of biological perturbations, the occurrence of post-translational modifications and the distribution of specific proteins within cells (Anderson & Anderson 1998).

Proteomics is recognized as an extremely important tool in the study of many biological systems. However, to date there has been only limited application of proteomics to address questions in marine ecological research. In the present study, a brief introduction to

*Email: jllopez@usc.es

© Inter-Research 2007 · www.int-res.com

proteomics is provided and the advantages, disadvantages, and some of the limitations of the field are discussed. The potential for the use of proteomics to address questions in marine ecology is illustrated by our research activities on marine bivalves.

GENOMICS VS. PROTEOMICS

The field of genomics utilizes a variety of technologies to study the information content of cells, i.e. their DNA or RNA. However, the phenotype that the genotype yields is dependent on interactions amongst its genes, the metabolic chemistry of the organism (internal environment) and environmental factors (external environment). Understanding how physiological, environmental and ecological factors (and the time span over which they occur) affect the internal and external environment and ultimately the phenotype of organisms is critical for our understanding of many areas of marine ecology. Proteomics provides us with many necessary tools with which we can improve our understanding of these complex relationships.

The field of proteomics is complementary to genomics in that it provides additional information on gene expression and regulation. Proteomics also enables the analysis of other biological processes that lead to the production of proteins. For example, the analysis of transcription alone provides a limited view of gene expression because it does not take into account regulatory steps at the level of mRNA translation. The poor correlation between the amount of mRNA and their respective proteins in cells was first demonstrated by Anderson & Seilhammer (1997). mRNA is a disposable message that has no other function than to temporarily serve to convey a piece of information, whereas protein measurements relate directly to functional mechanisms. In addition, post-transcriptional changes such as alternative gene splicing and post-translational modifications of proteins, such as glycosylation or phosphorylation, significantly increase the number of different proteins above that predicted by DNA or mRNA analysis alone. With respect to post-translational modifications, it is known that the activity of proteins is regulated by their modification state. Therefore, it is possible that even though the expression of a gene may be the same in 2 situations, differences in the phosphorylation status may result in significant differences in the activity of the proteins produced. The use of transcriptomics alone provides only partial information on such changes.

Protein function and the phenotypic traits of a particular genotype depend not only on the proteins present and their possible post-translational modifications, but also on their levels of expression. The use of proteomics to measure changes in the levels of expression at protein level has enabled rapid advances in our understanding of the ecological and environmental adaptations of organisms, as well as of the biogeographical distribution of species (López et al. 2001, 2002, Fuentes et al. 2002). Proteomics provides a higher level of analysis to aid the understanding of gene function in particular and biology in general (López 2005).

PROTEOMICS TOOLS AND MARINE ECOLOGY

A comprehensive description of the proteome of an organism not only provides a catalogue of all proteins encoded by the genome but also data on protein expression under defined conditions (López 2005). For proteomics to be widely adopted, a robust technology must be established that allows the large-scale research needed for a holistic approach to protein science. A fundamental technology in proteomics studies is high-resolution 2-dimensional electrophoresis (2DE), which is a powerful technique used to separate complex mixtures of denatured proteins according to their charge and molecular weight (O'Farrell 1975). Combined with non-specific protein staining, the technique permits the visualization of a very large number of gene products that represent the more abundant proteins in a cell, tissue or organism. In addition, 2DE allows for the detection of some post- and co-translational modifications of proteins, which cannot be predicted from DNA sequences or transcriptomics (Anderson & Anderson 1998).

2DE has been used to generate large amounts of proteomics data for a wide variety of biological systems (Anderson & Anderson 1998, Jungblut et al. 1999, D'Ambrosio et al. 2005). In the marine environment, 2DE has been used to screen organisms for the presence of bioactive compounds and to detect and quantify changes in gene expression at the protein level during development, as well as in response to different physiological and environmental conditions (e.g. López et al. 2001, 2002, 2005, Olsson et al. 2004, Schweder et al. 2005, Barneah et al. 2006, McDonagh et al. 2006). 2DE also has great potential for the study of genetic variability of populations, in that it allows a more representative sample of the genome to be analyzed. However, studies of genetic variability in natural populations of animal species by means of 2DE are relatively scarce. This is because 2DE is technically more difficult and time-consuming than conventional 1-dimensional electrophoresis (1DE); furthermore, early studies that used 2DE revealed substantially less genetic variation than had been estimated by 1DE (Edwards & Hopkinson 1980, Aquadro & Avise 1981,

Neel 1990). Mosquera et al. (2003) successfully used 2DE to determine the degree of genetic variability for loci that encode abundant proteins in the marine mussel *Mytilus galloprovincialis*. In addition to demonstrating that 2DE can be used to study interpopulation genetic variability in *M. galloprovincialis*, Mosquera et al.'s (2003) study also compared the results obtained by 2DE and 1DE and discussed the possible reasons for the differences observed between these 2 approaches. This study was also the first to use 2DE in an attempt to detect linkage disequilibrium between loci that encode abundant proteins. Among a total of 406 two-locus pairs analyzed for the detection of linkage disequilibrium in the population sample, 92 showed statistically significant associations. Proteomics as a tool for genetic mapping needs to be further explored, especially because information on linkage of genetic markers in marine organisms is very scarce (Beumont 1994).

Since its initial development, 2DE has improved significantly. Improvements include: simplification and standardization of the methodology, the ability to load larger amounts of protein, thereby allowing the identification and analysis of less-abundant proteins, and better reproducibility between gels. These changes, along with reduced costs, now make it possible for more laboratories to take advantage of 2DE.

Other protein separation and quantification techniques include liquid chromatography (LC), high pressure liquid chromatrography (HPLC), and capillary electrophoresis (CE) (reviewed by Liebler 2002). Compared with 2DE, the amount of sample that can be used with multi-dimensional chromatography (LC/LC-MS/MS) is less restricted, the process is easier to automate, and specific classes of proteins such as very acidic, very basic, and membrane proteins—difficult to detect using 2DE—may be more readily detected. However, these techniques rely on digestion of the proteome into a complex peptide mixture before LC separation. It is questionable whether these techniques retain the ability to study proteolytic processing and post-translation modifications, which can be readily detected by 2DE. Nevertheless, a distinct advantage of the peptide-based techniques is the ability to perform very rigorous relative protein quantification between samples using isotopic labelling techniques such as iTRAQ.

Regardless of the technique used for separation, proteins are identified by mass spectrometry (MS) and bioinformatics analysis. MS allows protein identification and characterization with speed and accuracy (Aebersold 1993). It is mandatory for rapid proteomics development and plays a central role in proteome research today (Shevchenco et al. 1996). Several types of MS techniques can be used to identify proteins, e.g. peptide mass fingerprinting and partial sequencing by tandem MS, but a detailed review of these techniques is beyond the scope of this study. In addition to the quantification and identification of proteins, recent advances in MS now enable studies of post-translational changes in proteins (Figeys & Aebersold 1997, Carr et al. 2005).

An important consideration at the onset of any proteomics study is that the protein separation method must be able to produce polypeptides in a form that is compatible with the MS technique to be used. For example, if one selects the nano-electrospray ionisation principle (Mann & Wilm 1995), the sample prefractionation should terminate in a liquid form and should be separated at the end of the procedure by micro-LC, HPLC or CE. If one prefers matrix-assisted laser desorption ionisation (MALDI) (Patterson & Aebersold 1995), then the polypeptide of interest needs to be in a form that can be deposited on a solid target. For this and other reasons, researchers who wish to apply proteomics to their research are advised to consult with proteomics experts during the design of their studies. Fortunately, many universities and research centers now have laboratories or services that support proteomics research. These facilities have the knowledge and equipment necessary to deal with protein samples from diverse biological sources.

BIOINFORMATICS SUPPORT

Proteomics and genomics research generates large data sets, which must be organized, stored, and made accessible in logical ways. One of the key components of genome and proteome research is bioinformatics. Several categories of bioinformatic tools are required for proteome analysis (Haoudi & Bensmail 2006). Briefly these can be classified into those used for quantification and those used for protein sequence analysis. With respect to quantification, a variety of software-based image analysis tools are available to monitor and quantify proteins separated by 2DE or to facilitate the quantification of isotope-labelled peptides. Protein sequence analysis depends upon a variety of analytical tools in order to search databases for peptide and protein matches, as well as to predict structure and function.

High quality and well-annotated genomics and protein databases are the core of proteome research. In most instances, the characterization and identification of proteins by a proteomic approach is dependent on the existence of genomic resources for the organism of interest, or at least for closely related organisms. When working with samples from many marine species, we are limited by the availability of genomics and proteomics resources for those species and, in many instances, even closely related species. For example,

López et al. (2001) compared differences in protein expression between intertidal and cultured populations of *Mytilis galloprovincialis* using high resolution 2DE. Over 750 proteins that were consistently expressed in foot tissues were observed in that study. From these, 92 proteins were selected for additional analysis and statistically significant differences in protein abundance for almost 50% of these proteins were identified.

In another study, a proteomic approach was used to generate proteomics reference maps and subsequently to detect, quantify, and compare the global protein expression between 2 related species of marine mussel, *Mytilus edulis* and *M. galloprovincialis*, growing in their respective geographical habitats (López et al. 2002). A comparative study of the protein profiles generated from analytical 2DE gels was performed, and changes in protein expression were analyzed quantitatively by computer analysis. On average 1278 proteins were detected per gel and, of these, 420 proteins were selected for quantification. Of these, 15 proteins showed higher expression in *M. edulis* and 22 proteins in *M. galloprovincialis*. The technique of peptide mass fingerprinting using MALDI-TOF (matrix assisted laser desorption ionization-time of flight) and/or nanoelectrospray MS/MS was then applied to identify these differentially expressed proteins. We were able to unambiguously identify only 15 of these 37 proteins using these techniques. Our results demonstrated the sensitivity of 2DE when detecting differences in protein expression. However, our ability to only identify 41% of these differentially expressed proteins revealed an important limitation with respect to protein identification using peptide mass fingerprinting, which is that proteins can only be identified if their sequence (or a sequence of the same protein from a closely related species) is available for comparison. The poor characterization of *Mytilus* spp. and other mollusk species at both the gemonic and proteomic levels is responsible for our limited ability to identify these differentially expressed proteins. With the development of new analytical methods that enable de novo sequencing (nano-ESI [nanoelectrospray time of flight], Q-TOF [quadrupole time of flight] etc.), the application of proteomics to marine organisms will become more routine (López et al. 2002, 2005).

Protemonics also shows great promise with respect to the identification of protein markers that would allow for precise and rapid species identification (López et al. 2005). This would be especially beneficial for the identification of marine species that are difficult to identify using morphological characteristics or are ambiguous with respect to their taxonomic status (López et al. 2002, 2005). In previous studies, López et al. (2002, 2005) demonstrated how proteomics can be used for the routine identification of species-specific peptides. Although we cannot foresee how many proteins are, in general, needed for the identification of closely related organisms, the high throughput and speed of analysis of the modern MALDI-TOF mass spectrometers would allow the extension of this kind of comparative study to include hundreds and even thousands of proteins from a large number of individuals, making the identification of speciesspecific peptide markers highly likely. Studies of this sort have the advantage that no information from genomic or proteomic databases is needed. Once potential species-specific peptides are identified, lower throughput MS techniques such as nano-ESI-IT MS could be used to perform a more detailed characterization of these markers. HPLC-tandem MS, focused on the peptide markers, may then be used for a fast and highly accurate confirmation of specific identification. Since this last technique is strictly quantitative, it might also be used as a routine technique for species identification. In addition, this technique could aid in the development of antibodies against species-specific peptides, which would allow the identification of these peptides in crude tissue extracts. Furthermore, this procedure (described by López et al. 2002) is also suitable for phylogenetic studies.

FUTURE CHALLENGES IN MARINE PROTEOMICS

One of the principal challenges in proteomics is to achieve a level of understanding of protein expression, post-translational modification and interaction that is similar in scope to what genomics has provided us for genes. This is a more difficult task with proteins than with nucleic acids because genes are approximately equimolecular in genomic DNA, whereas proteins may span 7 or 8 orders of magnitude in terms of functional abundance within a cell type, i.e. a functioning protein may be much less concentrated than other functioning proteins. In addition, there is difficulty in resolving very hydrophobic, very basic, or very large proteins using current 2DE systems. In proteomics, important discoveries will be made through quantitative observations of a limited (but large) number of protein gene products once the protein database is rich enough.

In response to technical challenges, we are likely to see the emergence of fully automated 2DE systems. Furthermore, continued development of non-gel-based alternative technologies that use combinations of capillary electrophoresis or multidimensional-HPLC coupled to MS will make proteomics data acquisition even more routine and possibly cost effective.

The major obstacle in the application of proteomics to many fields is generally considered to be data analy-

sis. However, with regard to the marine environment, a lack of genomic and proteomic resources for species of interest are often the major obstacle. Although it is possible to use data from related species, there are relatively few marine organisms for which sufficient genomic or proteomic data exist. This situation will likely improve in the near future owing to significant reductions in the time required for, and costs associated with, large-scale sequencing and proteomics studies. Such changes will make it economically feasible to begin to study a wider variety of organisms, including those from marine environments. The many genome projects that are planned or underway for numerous marine species will provide genomic resources that will greatly improve our ability to apply proteomics to the study of marine ecology. As mentioned previously, the development of reliable software tools that allow for the identification of proteins not represented in databases will greatly accelerate the rate at which proteomics is applied in marine science.

LITERATURE CITED

Aebersold R (1993) Mass spectrometry of proteins and peptides in biotechnology. Curr Opin Biotechnol 4:412–419

Anderson L, Seilhamer (1997) A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18:533–537

Anderson NG, Anderson L (1982) The human protein index. Clin Chem 28:739–748

Anderson NG, Anderson L (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19:1835–1861

Aquadro CF, Avise JC (1981) Genetic divergence between rodent species assessed by using two-dimensional electrophoresis. Proc Natl Acad Sci USA 78:3784–3788

Barneah O, Benayahu Y, Weiss VM (2006) Comparative proteomics of symbiotic and aposymbiotic juvenile soft corals. Mar Biotechnol 8:11–16

Beaumont AR (1994) Linkage studies in *Mytilus edulis*, the mussel. Heredity 72:557–562

Carr SA, Annan RS, Huddleston MJ (2005) Mapping post-translational modifications of proteins by MS-based selective detection: application to phosphoproteomics. Methods Enzymol 405:82–115

D'Ambrosio C, Arena S, Talamo F, Ledda L, Renzone G, Ferrara L, Scaloni A (2005) Comparative proteomic analysis of mammalian animal tissues and body fluids: bovine proteome database. J Chromatogr B 815:157–168

Debouck C, Goodfellow PN (1999) DNA microarrays in drug discovery and development. Nature Genet 21:48–50

Edwards Y, Hopkinson DA (1980) Are abundant proteins less variable? Nature 284:511–512

Figeys D, Aebersold R (1997) High sensitivity identification of proteins by electrospray ionization tandem mass spectrometry: initial comparison between an ion trap mass spectrometer and a triple quadrupole mass spectrometer. Electrophoresis 18:360–368

Fuentes J, López JL, Mosquera E, Vázquez J, Villalba A, Álvarez G (2002) Growth, mortality, pathological condicions and protein expresion of *Mytilus edulis* and *M. galloprovincialis* cosses cultured in the Ría de Arousa (NW of Spain). Aquaculture 213:233–251

Haoudi A, Bensmail H (2006) Bioinformatics and data mining in proteomics. Expert Rev Proteomics 3:333–343

Jungblut PR, Zimny-Arndt U, Zeindl-Eberhart E, Stulik J and 8 others (1999) Proteomics in human disease: cancer, heart and and infectious diseases. Electrophoresis 8:1217–1242

Liebler DC (2002) Introduction to proteomics. Tools for a new biology. Humana Press, Totowa, NJ

López JL (2005) Role of proteomics in taxonomy: the *Mytilus* complex as a model of study. J Chromatogr B 815:261–274

López JL, Mosquera E, Fuentes J, Marina A, Vázquez J, Álvarez G (2001) Two-dimensional gel electrophoresis of *Mytilus galloprovincialis*: differences in protein expresion between intertidal and cultured mussels. Mar Ecol Prog Ser 224:149–156

López JL, Marina A, Vázquez J, Álvarez G (2002) A proteomic approach to the study of the marine mussels *Mytilus edulis* and *M. galloprovincialis*. Mar Biol 141:217–223

López JL, Lorenzo S, Fuentes J (2005) Proteomic approach to probe for larval proteins of the mussel *Mytilus galloprovincialis*. Mar Biotechnol 7:396–404

Mann M, Wilm M (1995) Electrospray mass spectrometry for protein characterization. Trends Biochem Sci 20:219–224

McDonagh B, Tyther R, Sheehan D (2006) Redox proteomics in the mussel, *Mytilus edulis*. Mar Environ Res 62: S101–S104

Mosquera E, López JL, Álvarez G (2003) Genetic variability of the marine mussel *Mytilus galloprovincialis* assessed using two-dimensional electrophoresis. Heredity 90: 432–442

Neel JV (1990) Average locus differences in mutability related to protein 'class': a hypotesis. Proc Natl Acad Sci USA 87:2062–2066

O'Farrell (1975) High reolution two-dimensional electrophoresis of proteins. J Biol Chem 250:4007–4021

Olsson B, Bradley BP, Gilek M, Reiner O, Shepard JL, Tedengren M (2004) Physiological and proteomic responses in *Mytilus edulis* exposed to PCBs and PHAs extracted from Baltic Sea sediments. Hydrobiologia 514:15–27

Patterson SD, Aebersold R (1995) Mass spectrometry approaches for the identification of gel-separated proteins. Electrophoresis 16:1791–1814

Schweder T, Lindequist U, Lalk M (2005) Screening for new metabolites from marine microorganisms. In: Le Gal Y, Ulber R (eds) Advances in biochemical engineering/ biotechnology, Vol 96. Springer, Berlin, p 1–23

Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F and 5 others (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from 2 dimensional gels. Proc Natl Acad Sci USA 93: 14440–14445

Wilkins MR, Sánchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, Williams KL (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13:19–50

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: April 17, 2006; Accepted: November 6, 2006
Proofs received from author(s): February 5, 2007

Vol. 332: 281–289, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Marine proteomics

Brook L. Nunn[1,2], Aaron T. Timperman[3,*]

[1]Medicinal Chemistry Department, University of Washington, Box 355350, Seattle, Washington 98155, USA
[2]Marine Chemistry Department, University of Otago, Box 46, Dunedin, New Zealand
[3]C. Eugene Bennett Department of Chemistry, West Virginia University, Prospect St, Morgantown, West Virginia 26509-6045, USA

ABSTRACT: A wealth of information is recorded in a protein's primary sequence, which can be used to determine its biological function and origin, and provide clues to the mechanisms of degradation. In contrast to DNA, proteins and their amino acid constituents have demonstrated a wide-spread presence outside the cell, preserved in the environment. In marine samples, proteins are present as mixtures from numerous sources in a salty, complex matrix at low concentrations. As a result of these factors, studies of this nitrogen-based component in the oceans have previously been limited to bulk elemental and amino acid analyses; these analyses were incapable of providing details regarding protein sequence, function and source information. Advances in biological mass spectrometry now allow for the analysis and characterization of the protein component from the marine environment. Proteomic mass spectrometry is a high-throughput analysis of protein mixtures that does not require any prior knowledge of the original protein structures in the mixture, making it an ideal technique for marine studies. Potential marine applications of proteomics include: analyzing organisms cultured under different nutrient conditions to examine cellular expression and adaptation, profiling the marine dissolved and particulate organic matter pools to determine source information and understand long-term carbon preservation, and verifying genomic findings with proteomic analyses to determine which genes are translated and to what extent the protein is expressed. Although some major advances in marine studies and mass spectrometry have been made, there remains a significant amount of methods development and community education before the full potential of proteomics is reached.

KEY WORDS: Seawater · Protein · Mass spectrometry · Genomics · Dissolved organic matter · DOM · Particulate organic matter · POM

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Many researchers in the marine community focus their attention on better understanding the cycling and transfer of carbon in the world's oceans. The political and social awareness of global warming is currently a driving force toward an improved understanding of this pool of carbon and how it interacts with the atmosphere and influences the world's climate. As a result, there is an increasing need for more accurate models of the cycling and transfer of carbon throughout the oceans. Historically, oceanographers examined the carbon pool at the elemental level, analyzing bulk carbon concentrations, later followed by monomeric molecular level analysis including carbohydrates (Hecky 1973, Lyons et al. 1979, Cowie & Hedges 1984), hydrocarbons (Nissenbaum et al. 1971, Prahl 1985), lignin (Prahl 1985), and amino acids (King 1974, Lee & Cronin 1982, Henrichs & Farrington 1987, Lee 1988). Through more detailed, non-destructive analyses we can gain additional information on the origin and fate of these organic molecules. Recent advances and applications in molecular-level analyses, such as mass spectrometry and nuclear magnetic resonance (NMR), are now being applied to marine samples to gain a better understanding of size distribution and structures of the original organic polymeric molecules present (e.g. Minor et al. 2003, Kujawinskia et al. 2004, Li et al. 2004, Aluwihare et al. 2005).

One of the remaining untapped reservoirs of information is locked up in molecules that are common to all life and also persist in the environment as discrete

*Corresponding author. Email: atimperm@wvu.edu

© Inter-Research 2007 · www.int-res.com

units: proteins. Unlike the previous polymers analyzed (e.g. carbohydrates or hydrocarbons), proteins are the result of a precise arrangement of their monomeric constituents — amino acids — where the composition and sequence can be specific to source organisms and/or cellular function. Proteins comprise most of a cell's machinery and have many important functions including structural integrity, energy transfer, and cellular death. Protein expression is an important indicator of cellular state and can provide information on the activation of various cellular pathways, while the survival of particular proteins in the marine environment (e.g. in the dissolved or sedimentary pool) can provide insight into mechanisms that control the degradation of organic matter.

Proteins and their precursors, amino acids, are widespread in a variety of marine environments at significant enough concentrations to be considered an important contributor to the carbon and nitrogen pools (Hedges 1991, Benner et al. 1992, Keil et al. 1994, McCarthy et al. 1998, Horiuchi et al. 2004) (Fig. 1). Older protein-identification technologies allowed for the isolation and sequencing of proteins on a protein-by-protein basis. Because most marine investigators that are interested in this component of carbon are not looking for a specific protein, but instead the identification of any and all proteins in the system, the technology was not compatible with their requirements. Proteomics is a high-throughput analysis for the rapid identification of known or unknown protein mixtures in complex systems. The emergence of proteomics will allow investigators to sequence and enumerate as many proteins as possible from the system, and determine if these proteins change as a response to stimuli or environmental condition. With the improvements in technology and advancements in proteomics, marine investigations will now be able to gain greater information by examining these C and N components at a higher molecular level. Throughout the present study we discuss several themes and questions that have been previously approached by marine investigators; however, with the exception of a few studies (Tanoue 1996, Powell et al. 2005), all prior investigations on the protein component in the ocean have been limited to the analysis of amino acids rather than peptides and proteins. The goals of the present study are to: (1) introduce proteomic mass spectrometry and clarify some common misconceptions of data interpretation; (2) introduce potential applications of proteomics in the marine field; and (3) provide some ideas on how to advance the community at the pace of the technology. Although this technique is in its infancy in the marine field, it has the ability to provide many clues to the sources and transformation of carbon in the oceans.

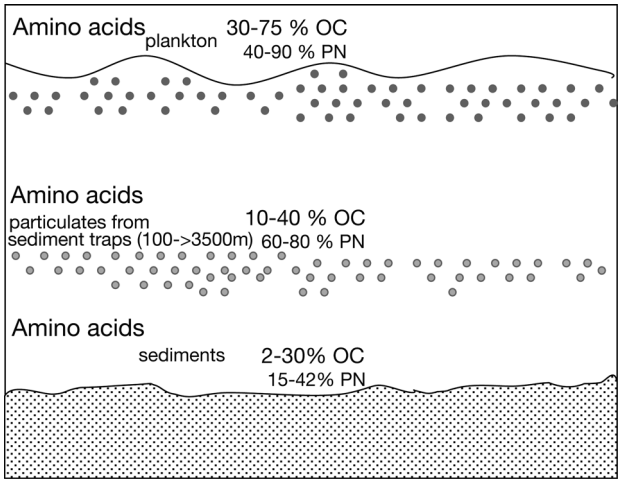

Fig. 1. Amino acids contribute a significant fraction of the percentage of organic carbon (OC) and particulate nitrogen (PN) in plankton (Siezen & Mague 1978, Lee & Cronin 1982, Lee & Olson 1984, Nguyen & Harvey 1994, 1997, Wakeham et al. 1997, Keil 1999 and references therein), particulates from sediment traps over a range of depths (Nguyen & Harvey 1994, Wakeham et al. 1997, Keil 1999 and references within), and in the coastal and deep ocean sediments (Wakeham et al. 1997, Keil 1999 and references therein, Nunn 2004). This, combined with other experimental evidence, strongly suggests that knowledge of the cycling and preservation of proteins in the marine environment is critical for understanding the global carbon cycle

METHODS OF PROTEIN ANALYSIS

Proteins are polymers consisting of a mixture of 20 genetically encoded amino acid monomers. The objective of protein analyses is to determine the order and number of amino acid residues that are covalently linked in a linear chain, referred to as the primary sequence. The primary sequence dictates how the protein is folded locally (secondary structure) and what form it takes 3-dimensionally (tertiary structure), which ultimately results in its biological role. The initial starting point for primary sequence analysis is to disrupt or denature its 3-dimensional structure, thereby unfolding the protein to make it more accessible for analysis. Previously, the majority of oceanic protein analyses involved complete hydrolysis of all peptide linkages, breaking proteins into the original amino acid monomers: complex mixtures of proteins and peptides were chemically hydrolyzed (150°C, 6 N HCl, 1 h) to amino acids for interpretation (e.g. Cowie & Hedges 1992, Keil & Kirchman 1993, McCarthy et al. 1997, Nunn & Keil 2005). As a result, any information that might have been gained pertaining to the sequence, structure, function or source of the protein was lost. Advances in biological mass spectrometry allow for mixtures of proteins to be analyzed from

more complex matrices and their primary sequences to be determined, thereby making the technique more informative to oceanographers.

Proteomic methods can be divided into techniques that analyze peptide fragments from the proteins, referred to as bottom-up protein analysis (Fig. 2), and those that analyze whole proteins in the mass spectrometer (MS), a top-down protein analysis (Reid & McLuckey 2002). In the bottom-up approach, proteins are cleaved into peptides to produce shorter segments

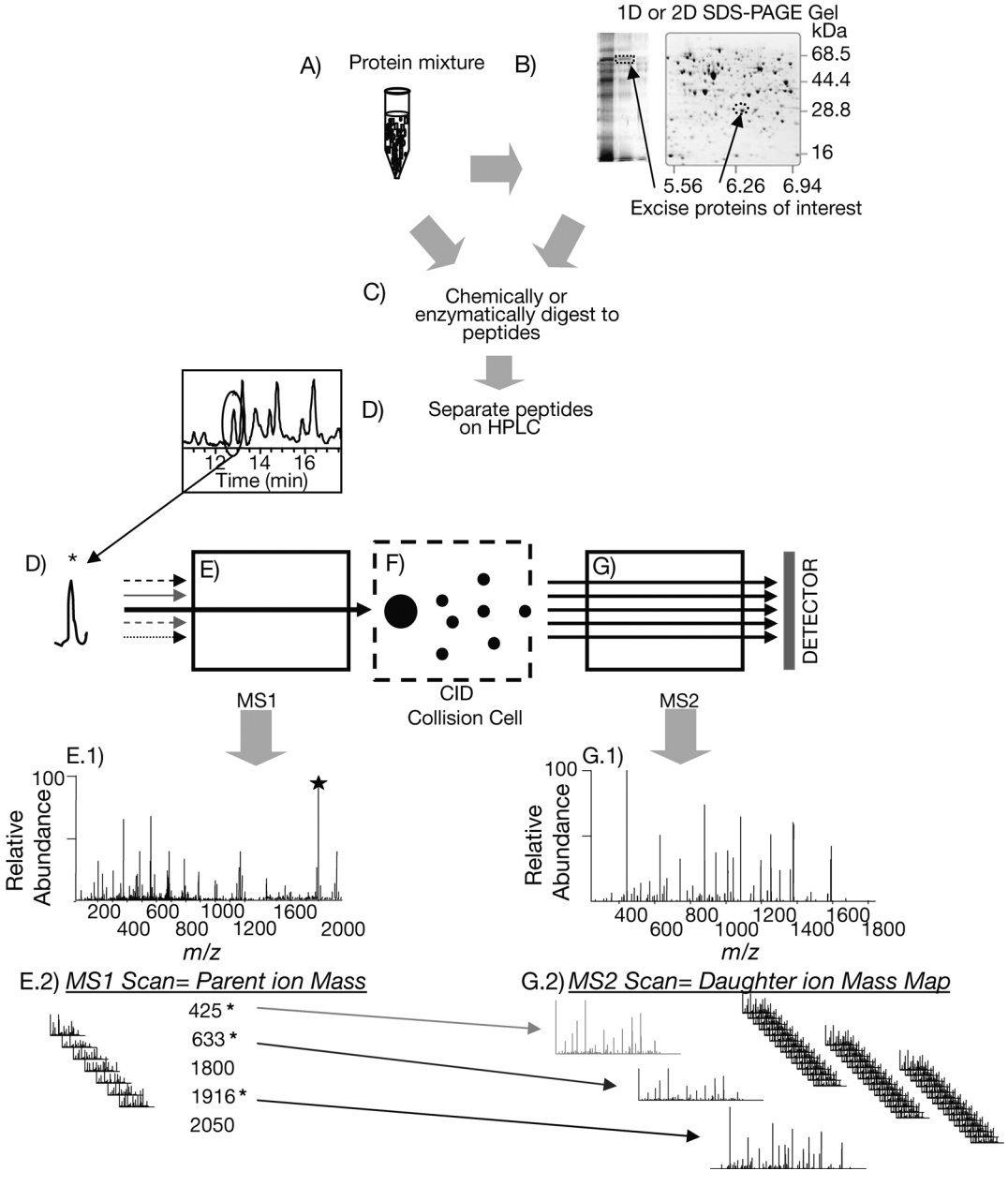

Fig. 2. Bottom-up proteomic project for protein identification from peptides using tandem mass spectrometry (MS). (A) Protein mixtures or (B) isolated proteins from gel electrophoresis can be sequenced using tandem MS approaches. The individual protein or mixture is first chemically or enzymatically digested into peptides (C). Complex mixture of peptides are then separated using inline HPLC (D) prior to injection and ionization in the mass spectrometer. As individual peptides (D) elute off the chromatography column they are ionized and analyzed in the first mass spectrometer (E: MS1). Mass to charge (m/z) ratios are measured (E.1), yielding the parent-ion scan. The analyst can then isolate single peptides (e.g. the 3 most intense peaks; E.2) from the parent-ion scan for fragmentation (F) and sequence determination. Each selected ion from the parent-ion scan is then individually fragmented and sent to the second MS (G: MS2), yielding daughter-ion scans (G.1). Sequence analysis is then performed using all parent-ion scans (E.2) and their respective daughter-ion scans (G.2). Interpretation of daughter-ion scans for the purpose of peptide sequencing is described in Fig. 3

that are more amenable to sequencing in the MS than whole proteins. Because peptides are unique to specific proteins, peptide tags or short peptide sequences that are determined experimentally can then be used to search the databases for the parent protein (e.g. Powell et al. 2005). Identification of more than 1 peptide unique to a protein is commonly used to infer the presence of the entire intact protein. As a result, bottom-up analyses excel at protein identification when combined with database searches. In contrast, the top-down approach analyzes whole proteins in the MS and can provide complete sequence coverage. The top-down method is therefore best suited for the analysis of protein modifications such as phosphorylations. Fragmentation of whole proteins (top-down) or peptides (bottom-up) can be achieved in the MS using one of a variety of dissociation technologies (e.g. electron capture dissociation). The present study focuses on the bottom-up approach of peptide sequencing and protein identification.

PEPTIDE SEQUENCING USING TANDEM MASS SPECTROMETRY

A basic knowledge of the fundamentals of peptide sequencing by tandem mass spectrometry is essential for understanding the potential applications for this technology; a more detailed description can be found in a number of recent publications (e.g. Fenn et al. 1989, Mann & Wilm 1994). Tandem mass spectrometry takes advantage of 3 properties of proteins: (1) the building blocks of proteins are known; (2) proteins can be cleaved into peptides; and (3) protonated peptides fragment in a predictable manner, producing product ion spectra that are reproducible and interpretable. The most commonly used proteomic method begins with the isolation of proteins using gel electrophoresis, followed by excision from the gel and proteolytic digestion of the protein using an enzyme, typically trypsin (Fig. 2A–C). The resulting peptides are then extracted and separated using reversed-phase high-performance liquid chromatography (HPLC, Fig. 2D), ionized, and the parent ion mass to charge (m/z) ratios are measured in the MS (Fig. 2E). An individual peptide parent ion can then be selected and isolated for fragmentation in the MS (Fig. 2F); the resulting m/z ratio values of the fragmented parent ions are measured, yielding a tandem mass spectrum (Fig. 2G). This ion isolation process is critical because it ensures that the fragment ions are from the selected parent ion, making this method extremely well suited to the analysis of complex mixtures.

In the positive ion mode, basic amino acid residues in the peptides are protonated. Frequently, tryptic peptides are doubly charged (+2) because both the amino-terminus (N-terminus) and the basic residue at the caboxy-terminus (C-terminus) are positively charged. The proton associated with the N-terminus in solution is mobile in the gas phase, allowing it to migrate along the peptide backbone and directing fragmentation to the adjacent amide bond. When fragmentation occurs at an amide bond, fragmentation ions that contain the N-terminal residue are called b-ions, whereas fragmentation ions that contain the C-terminal residue are referred to as y-ions. Fragmentation of a +2 parent ion typically results in a b- and y-ion that are each singly charged. Different members of the peptide ion population will typically break at different amide bonds, yielding b- and y-ion series (Fig. 3). The mass differences between singly charged ions that are contiguous in the series correspond to the amino acid residue masses; additionally, the residue order is encoded in the mass ladder (Fig. 3).

A single HPLC-MS run can produce thousands of spectra, making automated data filtering and interpretation a requirement (Hirosawa et al. 1993, Perkins et al. 1999). Automated analysis is typically achieved by comparing the experimentally obtained fragment ion spectra, with theoretical spectra mathematically predicted from the sequences in both genomic and protein databases. To perform correlative database sequence searching, the analyst typically provides the software with 3 pieces of information: the organism's full proteome (or genome for translation), the enzyme that was used for the digestion, and any chemical modifications or adducts that might be present (methylation, Na^+ adducts, etc.). Using scoring algorithms to rank the spectra, the software then returns a list of proteins, with their respective peptides identified, a final percent of protein sequenced, a correlation score, and HTML-links to the individual peptides' spectra for direct scrutiny. Since database correlation routines always return a match, proper filtering and manual verification are required to maintain reliability. For automated correlative database protein identifications, an important point of emphasis is that the interrogated protein or peptide sequence must be in the database in order for it to be properly identified. However, if the fragment-ion series is complete enough for a given peptide, the amino acid sequence can be mathematically interpreted directly from the tandem MS spectrum, either manually or using computerized algorithms (see Fig. 3). This ability to perform protein sequencing without depending on any prior knowledge of the amino acid sequence (de novo sequencing) is critical for environmental samples such as seawater, because only a small fraction of the contributing organisms' genomes have been sequenced (Powell et al. 2005). De novo sequencing programs are currently

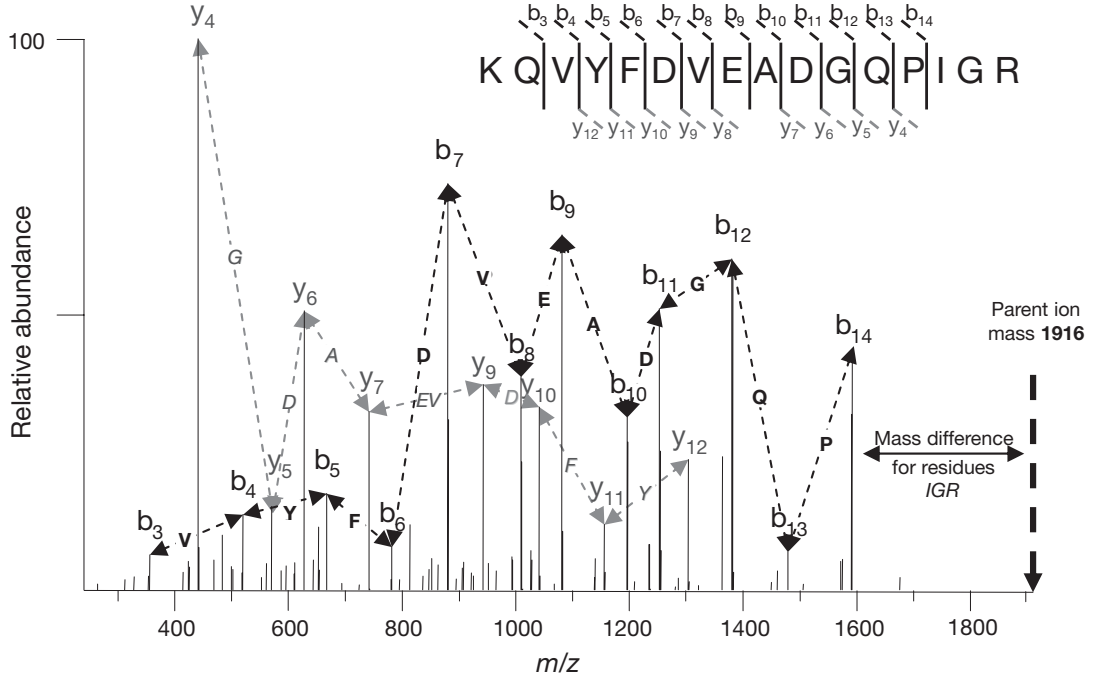

Fig. 3. How to de novo sequence a mass-to-charge daughter-ion spectra produced from the fragmentation of a tryptic peptide using tandem mass spectrometry (parent ion m/z = 1916 Da). Amino acid sequence of the original peptide denoted by amino acid single letter codes—amino-terminus is lysine (K), carboxy-terminus is arginine (R)—is indicated in the top-right corner. The b-ions (black) include all ions produced during fragmentation that retained the amino terminus (e.g. KQ, KQV, KQVY, etc.). The y-ion series (grey) consist of all ions produced during fragmentation that retained the carboxy-terminus (e.g. RGIP, RGIP, RGIPQ, etc.). The daughter-ion spectra is a plot of the m/z ratio of each ion produced during the fragmentation of the parent peptide and their relative abundance (y-axis). Mass differences between ions in the spectra are indicative of monoisotopic amino acid residue masses (e.g. b_{14} to b_{13} = 97 Da: Proline-**P**; b_{13} to b_{12} = 128 Da: Glutamine-**Q**). Dashed black lines indicate mass differences between b-ion series with the respective amino acid in the center of the line, and grey dashed lines with arrows are the mass differences between the y-ion series with their respective amino acids in the center of the line. Not all ions in both series need to be present to decipher the original amino acid sequence of the peptide

available and use scoring routines with similar caveats to database correlation programs.

For either approach it is important to note that the quality of the data is a major factor in the reliability and confidence of the sequences obtained. All aspects of sample preparation, separation and MS analysis can affect the data quality. Mass spectrometers that provide higher mass accuracy, resolving power, and signal-to-noise ratios produce higher quality data that will provide greater data reliability.

APPLICATIONS FOR MARINE PROTEOMICS

Current work in marine proteomics can be divided into 2 broad areas: the recovery and analysis of proteins from the marine environment, and the analysis of proteins from cultured organisms. Characterization of proteins directly from seawater can be used to determine the structures of proteins that are resistant to degradation and accumulate to detectable levels. The isolation and characterization of proteins from marine samples, such as seawater, porewater, particulates, and sediments, will greatly improve our understanding of the sources and mechanisms that control the cycling and long-term preservation of organic matter. For years the marine community has been limited to the examination of amino acids; full characterization of proteins from marine samples will provide a description of dissolved organic matter (DOM) components at the molecular level. Through the sequencing of these proteins and peptides, we can potentially gain information about the presence or past existence of an organism in a sample and the original function of the protein. We can also identify specific protein families, domains or themes preferentially preserved or any chemical modifications or adductions that might have enhanced the proteins' preservation. Through the combination of all these analyses we can greatly exceed previous elemental-level investigations by providing clues to what environmental conditions might encourage or discourage long-term preservation of carbon and nitrogen within the ocean. In a more directed strategy, specific proteins have been injected

into sediments and their degradation followed as a function of time (Nunn et al. 2003). These studies have shown that the model proteins used degrade rapidly, in the order of weeks.

Cultures of marine plankton can be used to determine what proteins from their genome are expressed and the relative levels at which these proteins are excreted, or released, into the surrounding environment. Since an organism's expressed proteome is dynamic, cellular protein expression changes as a function of environmental conditions. As a result, proteomics allows investigators to determine how organisms are able to biochemically cope and respond to varying environmental stresses. For example, proteins excreted from an organism into the surrounding medium could have one of numerous functions, including organism-to-organism communication or signaling (e.g. Wisniewska et al. 2003), or as an aid in the digestion or acquisition of nutrients, or as a microbial deterrent (e.g. Thomas et al. 2004). Isolating and sequencing these excreted proteins can inform us as to how the organisms biochemically manage and respond to their surroundings. In many parts of the world's oceans, different nutrients are in high demand as a result of being present at very dilute concentrations. A wide variety of organisms have adapted to these nutrient-deplete conditions and grow opportunistically when conditions are favorable. Thus, a long-standing question in the marine community concerns how these organisms are able to sequester the required nutrients from such dilute conditions. In many cases these questions can be answered using differential quantitative proteomics on organisms grown in culture with and without specific nutrients.

Controlled studies of cultured marine organisms can also improve our understanding of which peptide-linked molecules are most likely to contribute to the dissolved and particulate organic matter pools. Both relative protein expression levels and resulting protein products after extensive degradation can be analyzed and potentially quantified. Studies such as these may also provide information on relative resistance of different proteins to degradation, allowing for their selective enrichment and providing clues on long-term preservation (e.g. Nunn et al. 2003, Squier & Harvey 2006). Further insight into which degradation processes are most important may also be gained by controlled exposure of proteins to different enzymes, bacteria, light, or abiotic reactants. Using protein mass spectrometry, sequences and relative quantities of resulting peptide end-products can be obtained (Nunn et al. 2003, Peers & Price 2006). These types of experiments can provide the foundation for understanding which protein components comprise the recalcitrant dissolved and particulate organic matter pools in the ocean.

USING PROTEOMICS TO COMPLEMENT GENOMIC FINDINGS IN MARINE ECOSYSTEMS

In the past decade, ocean-based genomics has begun to explore the diversity, cellular evolution and adaptive abilities of marine organisms. Although this has provided the community with the beginnings of a database of microbial and eukaryotic blueprints, it does not necessarily translate into biochemical expression or phenotype. Genomics demonstrates which genes are shared, but proteomics can show clearer relationships by illustrating functional similarities and phenotypic variances. Through the use of pure genome sequences, open reading frames (ORFs) can be predicted, but they cannot be used to determine if or when transcription takes place or to what degree a protein is expressed. Proteomics can provide the researcher with more than the hypothetical cellular scenario. With a well-designed experiment, investigators can examine the conditions under which a protein is expressed (Nilsson & Davidson 2000, Kislinger & Emili 2003), its cellular location (Dunkley et al. 2004), the relative quantities (Yao et al. 2001, Molloy et al. 2005), and what protein–protein interactions take place (Giot et al. 2003, Schweitzer et al. 2003).

Because the ocean is one of the most dynamic environments in which organisms live, the success of a species depends on its ability to rapidly adapt to varying light, temperature and nutrient sources. Close examination of the genomes of oceanic microbes has already demonstrated that many of these organisms have the blueprints for diverse suites of organic and inorganic nitrogen and carbon transporters (Palenik et al. 2003, Armbrust et al. 2004). Proteomics can clarify if and to what extent various pathways are utilized, which environmental triggers act on the system, and relative protein-level response times. Additional information on protein expression levels in combination with gene expression will help investigators to clarify phylogenetic roots and possibly endosymbiotic events by highlighting dormant pseudo-genes, protein-level amino acid migrations, and mutations (Coin & Durbin 2004, Jaffe et al. 2004, Wirth et al. 2005). Through the use of proteomics, we may be able to simplify ocean-wide genomic investigations that are attempting to decipher evolutionary changes from ancestral cells. For example, instead of a broad-based survey of oceanic genomes, we can narrow the focus to a few directed analyses of proteins involved in specific biochemical pathways (Bibby et al. 2001, Strzepek & Harrison 2004, Peers & Price 2006).

THE FUTURE OF MARINE PROTEOMICS

If proteomic technology is beyond its tenth year (Wasinger et al. 1995), why is it that the marine field is

only recently beginning to use it as a tool to answer some of the community's questions? For many environmental investigators, molecular-level analyses have been impractical. The 3 primary reasons why the marine science field has taken so long to adopt the new technology are instrument availability, financial resources, and availability of trained personnel.

Excluding the proteomic investigation of cultured marine organisms, environmental protein analysts must contend with mixtures of proteins present at very low concentrations combined with complex matrices and relatively unknown sources. Prior to the recent investigation where large volumes (~100 l) of water were ultrafiltered to permit mass spectrometric characterization of the dissolved proteins (Powell et al. 2005), all previous investigations of this C and N pool involved amino acid hydrolysis and derivitization (Wakeham & Lee 1989, Cowie & Hedges 1992, McCarthy et al. 1997) as a means to circumvent low analyte concentrations present in high levels of contaminants. As proteomic technology is quickly being adopted by a number of different laboratories to investigate a wide variety of biological questions, rapid innovations and advances are being made to improve detection limits, sensitivity, and contaminant tolerance.

General improvements in the proteomics field are taking place, but because marine applications are in their infancy and it is such a specialized niche, there must first be substantial advances in the development of new methods. To analyze dissolved, exuded, preserved or particulate protein fractions from the ocean, samples must be collected (e.g. using sediment traps, large volumes of water, cultured organisms), extracted (e.g. chemically), de-contaminated (e.g. via chromatography, other chemical separation), isolated, or concentrated (e.g. via ultrafiltration, chemical precipitation, dialysis). MS techniques and instrumentation must then be optimized and a rigorous method for data analysis must be developed (e.g. de novo analysis, sequence homology searches) and validated (i.e. molecular weight or isoelectric point verification, immunoassays, or MS identification of synthetic peptides). To date, one of the primary limits for large-scale proteomic analyses is the lack of a marine genomic or proteomic database to search. In short, to finalize organism-level proteomic projects, there is a need for complete marine genomes. Several authors have addressed the complexity of this task because of the difficulty in isolating and culturing marine microbes (Beja 2004, Falkowski & de Vargas 2004, Hess 2004, Venter et al. 2004). Another obstacle that must be overcome before the completion of marine proteomics projects is the lack of facilities dedicated to large environmental protein discovery projects (not medical use). Typically only small projects are tackled as 'pet projects' by proteomic facilities and investigators, and often there is neither sufficient time nor instrumental resources to adequately develop techniques and identify marine proteins. This situation strongly implies the need to encourage funding for larger collaborative groups that include investigators not typically involved in the marine or oceanographic community.

In order to investigate some of the larger marine proteomics questions or to complement marine genomes with proteomes, funding for environmental research will need to increase. An efficient proteomics facility typically requires several qualified, full-time technical staff to work together as a team to complete full annotations. The technical support includes people trained in wet-laboratory chemical preparations, protein chromatography, methods development, and instrumental optimization and maintenance, in addition to IT staff. Unlike the genomics field, there are to date no large proteomics facilities partially dedicated to helping answer environmental questions (e.g. Joint Genome Institute, California, USA). Focusing the funding on a few environmental proteomic centers may alleviate this problem and allow marine investigators to continue to explore and collect ancillary data from all over the world's oceans. The available funding also plays a role in the skilled personnel available for completing marine proteomics-based projects. Many students, doctorates, or staff trained in proteomic mass spectrometry can be easily enticed to migrate to the life sciences divisions where funding is higher, jobs are more prevalent, and resources are seemingly unlimited relative to environmental research. In order for marine proteomics to flourish, trained personnel will need to be recruited, and an awareness of the importance of solving global environmental questions must become a priority for both government and community.

CONCLUSIONS

Moving beyond the analysis of elemental concentrations and amino acids is the next step toward advancing the science of marine organic chemistry. Because proteins are an intricate arrangement of 20 amino acids, each one can be specific to both a function and a source. Recent advancements in the field of biological mass spectrometry now provide an avenue through which to analyze the proteomics of different marine systems. A recent study by Powell et al. (2005) demonstrated how this high-throughput analysis allowed them to investigate the DOM pool without the need for specialized techniques that only identify expected targets (e.g. enzyme assays, antibody assays, fluorescent tags). As a discovery driven science, proteomics allows users to identify complete unknowns without missing

unanticipated interactions. This dramatically improves the range of applications within the marine field for which this technique can be employed. However, because the marine field consists of such diverse environments and matrices in which these proteins reside (e.g. phytoplankton, sediment, hydrothermal vents), a great amount of methods development remains to be completed.

After sufficient methods development and general cataloguing of marine proteomes has occurred, biogeochemists will better be able to model the evolution and cycling of carbon pools within the ocean. We can begin to survey how different marine organisms' proteomes adapt to dynamic nutrient conditions, and which proteins are expressed in the cell, released into the environment, and passed between trophic levels. This information will provide great insight into which proteins are preserved in the environment and whether chemical modifications play a role in their ultimate preservation. The culmination of numerous marine proteomic studies has the potential to allow global-carbon investigators to model how marine organisms will respond to future anthropogenic perturbations and release proteins into the environment for long-term preservation. Integrating these techniques into the marine field is the next logical step to advancing oceanic environmental research.

Acknowledgements. We thank P. Boyd, C. Nunn, B. Gallis, and reviewers for their comments and suggestions for improving this manuscript. B.L.N. thanks the Office of Polar Programs for her NSF postdoctoral fellowship. This work was supported by National Science Foundation grant numbers 0444148 and OCE-0453737.

LITERATURE CITED

Aluwihare LI, Repeta DJ, Pantoja S, Johnson CG (2005) Two chemically distinct pools of organic nitrogen accumulate in the ocean. Science 308:1007–1010

Armbrust EV, Berges JA, Bowler C, Green BR and 41 others (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306: 79–86

Beja O (2004) To BAC or not to BAC: marine ecogenomics. Curr Opin Biotechnol 15:187–190

Benner R, Pakulski JD, McCarthy M, Hedges JI, Hatcher PG (1992) Bulk chemical characteristics of dissolved organic matter in the ocean. Science 255:1561–1564

Bibby TS, Nield J, Barber J (2001) Iron deficiency induces the formation of an antenna ring around trimeric photosystem I in cyanobacteria. Nature 412:743–745

Coin L, Durbin R (2004) Improved techniques for the identification of pseudogenes. Bioinformatics 20(Suppl 1): I94–I100

Cowie GL, Hedges JI (1984) Determination of neutral sugars in plankton, sediments, and wood by capillary gas chromatography of equilibrated isomeric mixtures. Anal Chem 56:497–504

Cowie GL, Hedges JI (1992) Sources and reactivities of amino acids in a coastal marine environment. Limnol Oceanogr 37:703–724

Dunkley TPJ, Dupree P, Watson RB, Lilley KS (2004) The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in *Arabidopsis thaliana*. Biochem Soc Trans 32: 520–523

Falkowski PG, de Vargas C (2004) Genomics and evolution. Shotgun sequencing in the sea: a blast from the past? Science 304:58–60

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Analytical instrumentation: electrospray ionization for mass spectrometry of large biomolecules. Science 246: 64–71

Giot L, Bader JS, Brouwer C, Chaudhuri A and 45 others (2003) A protein interaction map of *Drosophila melanogaster*. Science 302:1727–1736

Hecky RE, Mopper K, Kilham P, Degens ET (1973) The amino acid and sugar composition of diatom cell-walls. Mar Biol 19:323–331

Hedges J (1991) Lignin, cutin, amino acids and carbohydrate analyses of marine particulate organic matter. In: Hurd DC, Spencer DW (eds) Marine particles: analysis and characterization. American Geophysical Union, Washington, DC, p 129–137

Henrichs SM, Farrington J (1987) Early diagenesis of amino acids and organic matter in two coastal marine sediments. Geochim Cosmochim Acta 51:1–15

Hess WR (2004) Genome analysis of marine photosynthetic microbes and their global role. Curr Opin Biotechnol 15: 191–198

Hirosawa M, Hoshida M, Ishikawa M, Toya T (1993) Mascot–multiple alignment system for protein sequences based on 3-way dynamic-programming. Comp Appl Biosci 9: 161–167

Horiuchi T, Takano Y, Ishibashi J, Marumo K, Kobayashi K (2004) Amino acids in water samples from deep sea hydrothermal vents at Suiyo Seamount, Izu-Bonin Arc, Pacific Ocean. Org Geochem 35:1121–1128

Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4:59–77

Keil RG (1999) Early diagenesis of amino acids in high organic content marine sediments. Geochem Earth Surf 5:259–262

Keil RG, Kirchman DL (1993) Dissolved combined amino acids: chemical form and utilization by marine bacteria. Limnol Oceanogr 38:1256–1270

Keil RG, Montluçon DB, Prahl FG, Hedges JI (1994) Sorptive preservation of labile organic matter in marine sediments. Nature 370:549–551

King KJ (1974) Preserved amino acids from silicified protein in fossil Radiolaria. Nature 252:690–692

Kislinger T, Emili A (2003) Going global: protein expression profiling using shotgun mass spectrometry. Curr Opin Mol Theor 5:285–293

Kujawinskia EB, Vecchio RD, Blough NV, Klein GC, Marshall AG (2004) Probing molecular-level transformations of dissolved organic matter: insights on photochemical degradation and protozoan modification of DOM from electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Mar Chem 92:23–37

Lee C (1988) Amino acid chemistry and amine biogeochemistry in particulate material and sediments. In: Blackburn TH, Sorensen J (eds) Nitrogen cycling in coastal marine environments. SCOPE Series 33. John Wiley & Sons, p 125–141

Lee C, Cronin C (1982) The vertical flux of particulate nitrogen in the sea: decomposition of amino acids in the Peru upwelling area and the equitorial Pacific. J Mar Res 40: 227–251

Lee C, Olson BL (1984) Dissolved, exchangeable and bound aliphatic amines in marine sediments: initial results. Org Geochem 6:259–263

Li L, Zhao Z, Huang W, Peng P, Sheng G, Fua J (2004) Characterization of humic acids fractionated by ultrafiltration. Org Geochem 35:1025–1037

Lyons WB, Gaudette HE, Hewitt AD (1979) Dissolved organic matter in pore water of carbonate sediments from Bermuda. Geochim Cosmochim Acta 43:433–437

Mann M, Wilm M (1994) Error tolerant dentification of peptides in sequence databases by Peptide Sequence Tags. Anal Chem 66:4390–4399

McCarthy M, Pratum T, Hedges J, Benner R (1997) Chemical composition of dissolved organic nitrogen in the ocean. Nature 390:150–154

McCarthy MD, Hedges JI, Benner R (1998) Major bacterial contribution to marine dissolved organic nitrogen. Science 281:231–234

Minor EC, Wakeham SG, Lee C (2003) Changes in the molecular-level characteristics of sinking marine particles with water column depth. Geochim Cosmochim Acta 67: 4277–4288

Molloy MP, Donohoe S, Brzezinski EE, Kilby GW, Stevenson TI, Baker JD, Goodlett DR, Gage DA (2005) Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling. Proteomics 5: 1204–1208

Nguyen RT, Harvey HR (1994) A rapid micro-scale method for the extraction and analysis of protein in marine samples. Mar Chem 45:1–14

Nguyen RT, Harvey HR (1997) Protein and amino acid cycling during phytoplankton decomposition in oxic and anoxic waters. Org Geochem 27:115–128

Nilsson CL, Davidson P (2000) New separation tools for comprehensive studies of protein expression by mass spectrometry. Mass Spectr Rev 19:390–397

Nissenbaum A, Baedecker MJ, Kaplan IR (1971) Studies on dissolved organic matter from interstitial water of a reducing marine fjord. In: Gaentner HRV, Wehner H (eds) Advances in organic geochemistry. Pergamon Press, New York, p 427–440

Nunn BL (2004) Moving beyond amino acids: examinations of the protein component in marine sediments. PhD dissertation, University of Washington, Seattle

Nunn BL, Keil RG (2005) Size distribution and chemistry of proteins in Washington coast sediments. Biogeochemistry 75:177–200

Nunn BL, Norbeck A, Keil RG (2003) Hydrolysis patterns and the production of peptide intermediates during protein degradation in marine systems. Mar Chem 83:59–73

Palenik B, Brahamsha B, Larimer FW, Land M and 11 others (2003) The genome of a motile marine *Synechococcus*. Nature 424:1037–1042

Peers G, Price NM (2006) Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature 441:341–344

Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Powell MJ, Sutton JN, Del Castillo CE, Timperman AI (2005) Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. Mar Chem 95:183–198

Prahl FG (1985) Chemical evidence of differential particle dispersal in the southern Washington coastal environment. Geochim Cosmochim Acta 49:2533–2539

Reid GE, McLuckey SA (2002) 'Top down' protein characterization via tandem mass spectrometry. J Mass Spectr 37: 663–675

Schweitzer B, Predki P, Snyder M (2003) Microarrays to characterize protein interactions on a whole-proteome scale. Proteomics 3:2190–2199

Siezen RJ, Mague TH (1978) Amino acids in suspended particulate matter from oceanic and coastal waters of the Pacific. Mar Chem 6:215–231

Squier AH, Harvey HR (2006) Applying proteomics to geochemistry: tracking individual proteins during marine diatom degradation using liquid chromatography-tandem mass spectrometry. EOS Trans: AGU Ocean Sci Meet Suppl 87, Abstr OS25L-01

Strzepek RF, Harrison PJ (2004) Photosynthetic architecture differs in coastal and oceanic diatoms. Nature 431: 689–692

Tanoue E (1996) Characterization of the particulate protein in Pacific surface waters. J Mar Res 54:967–990

Thomas X, Destoumieux-Garzon D, Peduzzi J, Afonso C and 7 others (2004) Siderophore peptide, a new type of post-translationally modified antibacterial peptide with potent activity. J Biol Chem 279:28233–28242

Venter JC, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Wakeham SG, Lee C (1989) Organic geochemistry of particulate matter in the ocean: the role of particulates in oceanic sedimentary cycling. Org Geochem 14:83–96

Wakeham SG, Lee C, Hedges J, Hernes PJ, Peterson ML (1997) Molecular indicators of diagenetic status in marine organic matter. Geochim Cosmochim Acta 61:5363–5369

Wasinger VC, Cordwell SJ, Cerpa-Poljak A, Yan JX and 6 others (1995) Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. Electrophoresis 16: 1090–1094

Wirth B, Louis VL, Potier S, Souciet JL, Despons L (2005) Paleogenomics or the search for remnant duplicated copies of the yeast DUP240 gene family in intergenic areas. Mol Biol Evol 22:1764–1771

Wisniewska J, Trejgell A, Tretyn A (2003) Plant signaling peptides. Acta Physiol Plant 25:105–122

Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal Chem 73:2836–2842

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: July 19, 2006; Accepted: November 8, 2006
Proofs received from author(s): February 20, 2007

Vol. 332: 291–299, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Integration of genomics and proteomics into marine microbial ecology

Torsten Thomas, Suhelen Egan, Dominic Burg, Charmaine Ng, Lily Ting, Ricardo Cavicchioli*

School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia

ABSTRACT: Genomics and proteomics of microorganisms are revolutionizing our understanding of marine microbial ecology. In this essay we address this by discussing (1) what microbial genome resources are available for marine ecologists, (2) how single-organism genomics and proteomics have revealed new microbial functions in the marine ecosystem, and (3) how the integration of metagenomics, metaproteomics and biogeochemical studies will further advance the field of marine microbial ecology. Comprehensive knowledge of the genetic blueprints, the functions and the interactions of microbial communities will provide insight into the evolution of marine ecosystems and enable rational predictions of how microbial processes will affect, and be affected by, environmental changes.

KEY WORDS: Environmental genomics · Metagenomics · Metaproteomics · Proteomics · Marine ecology · Marine microbiology

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Marine microbial ecology has advanced over the last decade through a progression of approaches that has included taxonomic and physiological studies of culturable isolates, molecular community analysis (e.g. rRNA), analyses of complete genome sequences of individual isolates and, most recently, metagenomic analyses of entire microbial communities. Concomitant with technological advances of the genomic era has been an exponential increase in the extent of data pertaining to marine microorganisms. This has provided on the one hand an enormous capacity to learn, while on the other a daunting overload of information. It is clear that irrespective of the quantity, information lacks real value unless the intelligent means are available to process it effectively. This essay reflects on how genomics and proteomics may empower marine ecological studies. It considers strengths and difficulties of marine genomics and proteomics, and discusses the need to integrate these data with the full gamut of all available data (e.g. physical, geochemical) that describe the marine system. It is apparent that getting the most out of the genome stockpile will require healthy and informed interplay among scientists in many disciplines.

MICROBIAL GENOME RESOURCES FOR MARINE ECOLOGISTS

In the last 10 years, microbial genomics has experienced one of the most dramatic developments and advances of any scientific field. Microbiologists are now facing a genomic 'data flood' with more than 300 finished, bacterial or archaeal genome sequences available and over 900 more in progress (Liolios et al. 2006). Improved sequencing technologies and strategies (Margulies et al. 2005, Goldberg et al. 2006, Zhang et al. 2006) will continue to support this trend. In the near future, genome sequencing of new microorganisms will become a standard tool for microbial characterisation, analogous to the use of Gram staining in the past. Large-scale sequencing programs, such as

*Corresponding author. Email: r.cavicchioli@unsw.edu.au

© Inter-Research 2007 · www.int-res.com

the Microbial Genome Sequencing Project of the Betty and Gordon Moore Foundation, are already involved in the sequencing of numerous (>100) marine microbial species (www.moore.org/microgenome). Several of these organisms have been isolated in global diversity studies, and genome sequencing will now fast-track the understanding of their biology.

There would be few scientists who would argue against the immense value of the growing number of marine microbial genome sequences, and most microbiologists engaged in laboratory-based physiological or evolutionary studies will have effectively integrated available genome-based knowledge into their research. However, the integration of microbial genomics into marine ecology has not been as rapidly or as widely adopted. This may be due in part to the traditional background training of marine ecologists, which has focused less on the properties of individual organisms and more on the broader properties of the ecosystem, and to the lack of ecological data linked to genome sequence data of marine microorganisms (see last paragraph of this section). Some of this can be remedied immediately by simply tapping into available web-based resources (e.g. becoming familiar with what type of information is available and beginning to find out fundamental information about target organisms). Comprehensive databases and user-friendly, web-based interfaces have made genomic information increasingly accessible, without the need for specialized bioinformatics training or knowledge. The Integrated Microbial Genome (IMG) database of the Joint Genome Institute (JGI) and the Comprehensive Microbial Resource (CMR) of The Institute of Genomic Research (TIGR) are just 2 examples of the excellent tools available that allow the user to view, browse, analyse and compare microbial genome information. Specialized interest groups also provide databases dedicated to particular microbial groups, such as the Roseobase (http://roseobase.org/), which deals with genomic information of the abundant, marine Roseobacter clade. Table 1 lists some databases and web-based tools relevant to marine microbial genomics.

There are a number of issues in genomics that broadly affect the genomics community and that haven't been resolved, and there are additional aspects that need to be addressed in order to effectively facilitate ecological studies. Maintaining data quality is a broadly important issue with genome sequence data. The sheer volume of DNA sequence data in combination with limited human resources has made it extremely difficult to carefully and manually revise and curate the data. This has resulted in genome sequences being wrongly assembled from raw data (Salzberg & Yorke 2005) and automated gene prediction or annotation processes being inaccurate (Nielsen & Krogh 2005). Database users should therefore be cautious with predicted genome properties (particularly from auto-annotation pipelines) and be aware of the need to critically review the evidence for assigned gene function. For example, if a gene has been annotated based on experimental evidence or high similarity to a gene that has been experimentally characterised, the functional prediction is likely to be sound. However, there are numerous examples of annotations

Table 1. Web resources for marine, microbial genomics

Name	Weblink	Description
Integrated Microbial Genomes	http://img.jgi.doe.gov	Comparative database, all publicly available genomes
Comprehensive Microbial Resource	http://cmr.tigr.org	Comparative database, all publicly available microbial genomes
Microbial Genome Database for Comparative Analysis	http://mbgd.genome.ad.jp	Comparative database, all publicly available microbial genomes
ERGO	http://ergo.integratedgenomics.com	Private, comprehensive database
Center for Biological Sequence Analysis	www.cbs.dtu.dk/index.shtml	Comprehensive database and several web-based tools
Megx.net	www.megx.net	Database resource for marine ecological genomics; in development
Camera	http://camera.calit2.net	Cyberinfrastructure for marine microbial ecology research and analysis
Roseobase	http://roseobase.org	Specialised database for marine *Roseobacter* strains
Cyanobase	www.kazusa.or.jp/cyano	Specialised database for cyanobacterial genomes
Moore Foundation Microbial Genome Sequencing Project	www.moore.org/microgenome	links to maps and genome database

(classification of gene function) linked with low or high similarity to a gene with poorly defined properties. For example, a number of genes in Archaea, likely to be aminopeptidase genes (Ando et al. 1999), have been annotated as cellulase genes. Owing to an initial mis-annotation in 1 genome, subsequent archaeal genes with high levels of similarity were consequently mis-annotated. As there is presently no easy solution to this widespread problem, it argues strongly for individuals to carefully examine important gene targets of interest and to enhance the level of functional analysis of genes in order to experimentally determine their functions.

Genome databases have mainly been designed to enhance understanding of the biology and evolution of individual organisms, and there is clearly a need to include information that is relevant to ecological research (Lombardot et al. 2006). Database fields that are lacking include information describing the physical habitat from where an organism has been isolated or is typically present (e.g. temperate or tropical waters, planktonic or surface-associated), additional biological information about the environment (e.g. competitors, viruses, predators), information about seasonal and spatial abundance of organisms in the community, and a summary of relevant oceanography and physico-chemical properties (e.g. O_2, minerals, salinity, dissolved/particulate organic carbon). This would allow macrobiotic or abiotic parameters to be linked with molecular or genomic properties, and hence provide a straightforward bridge between ecology and genomics. Engaging ecologists with database designers/curators would help to create more ecologically useful databases.

MARINE ECOLOGY IS ALREADY BENEFITING FROM MICROBIAL GENOMICS AND PROTEOMICS

Genome sequences

In an analagous manner to the analysis of the human genome to predict specific drug targets and candidates for gene therapy, genome sequencing of ecologically relevant microorganisms and microbial communities (metagenomics) can provide new insight or generate testable hypotheses about ecosystem function (see 'The way forward with marine microbial ecology is through an integrated 'meta' approach'). A striking example is the discovery of the light-dependent proton pump, bacterial proteorhodopsin, which was first discovered from the sequences of cloned environmental DNA (eDNA), and is thought to play a major role in the generation of energy for microbial metabolism in the oceans (Beja et al. 2000, 2001). Prior to this discovery, light-driven processes were mainly linked to processes such as primary production by photosynthetic cyanobacteria. Another good example is the prediction of archaeal-driven nitrification processes derived from the analysis of metagenomic data (Schleper et al. 2005, Hallam et al. 2006), and the verification of this ability through the isolation of a chemolithoautotrophic ammonia-oxidizing member of the Crenarchaeota (Konneke et al. 2005).

The sequencing of genomes of single microbial species is also clearly of value for deriving inferences about ecology of marine bacteria. For example, genomic studies of ubiquitous planktonic bacteria (the SAR11 isolate *Pelagibacter ubique*, and a member of the Roseobacter clade *Silicibacter pomeroyi*) have greatly enhanced our understanding of how some microorganisms have adapted and evolved to become numerically abundant within the marine environment (Moran et al. 2004, Giovannoni et al. 2005). *P. ubique* has the smallest known genome (1.3 Mb) of any free-living microorganism, and points to an evolutionary adaptive strategy involving genome streamlining; i.e. optimizing growth efficiency by minimizing the genomic and cellular complement that needs to be reproduced in order for the species to survive and remain evolutionarily competitive (Giovannoni et al. 2005). Despite the relatively small size of the genome, *P. ubique* still possesses the capacity to synthesise all 20 amino acids and all core functions required for a free-living bacterium. The small genome size appears to have been selected through a process leading to the minimization of non-functional DNA, extra-chromosomally derived genetic elements (e.g. phage, integrons or transposons), and duplicated genes. Comparative studies with genome sequences of species from the same ecosystem indicate that adaptation to oligotrophy in this organism involves a low level of gene regulation and an investment in genes devoted to energy metabolism and high-affinity nutrient uptake.

Silicibacter pomeroyi is a dominant member of the coastal bacterioplankton (Moran et al. 2004). Based on genome sequence analysis, it appears that *S. pomeroyi* takes an opportunistic strategy towards nutrient acquisition. Genes for cell-density-dependent regulation, rapid growth and uptake systems for algal-derived compounds are present, suggesting that the organism is capable of associating with nutrient-rich hot-spots such as algal plankton and other suspended particles. Furthermore, the presence of gene clusters encoding enzymes for the oxidation of reduced inorganic compounds (e.g. carbon monoxide and sulphide) suggests that *S. pomeroyi* is a lithoheterotroph that gains energy from inorganic compounds and uses organic carbon compounds that are at low abundance for generating bacterial biomass. Arising from this genome

sequence analysis, a range of experimental studies can be designed and performed to assess the proposed metabolic capabilities and ecological function of *S. pomeroyi*. In view of the apparent wide-spread capacity for lithoheterotrophy that has been deduced from the analysis of metagenome data (e.g. the Sargasso Sea library; Venter et al. 2004) and the potential impact of this on global nutrient cycling in the oceans (Moran et al. 2004), performing functional studies to better understand this process is of considerable importance (see 'Functional genomics' below).

As great advances in marine ecology can come from the analysis of genome sequence data of individuals and communities of marine microorganisms, a strong argument can be made for a greater emphasis to be placed on the training of scientists with the necessary expertise in genomics, in order to more fully exploit the potential of genomic data for marine ecology research. In particular, there is an important need to develop synergies between bioinformaticians and ecologists/biologists, in order to translate the raw stock piles of genome sequence data into valuable science.

Functional genomics

Global functional studies, such as proteomics and transcriptomics, have the potential to most rapidly advance our understanding of functional cellular processes, and hence likely ecological processes. Studies relating to how an organism (or community) responds to environmental change provide insight into core physiological properties and adaptive strategies. Technological advances in the functional 'omics' have developed in concert with genome sequencing technology, particularly owing to the need for high-throughput procedures to keep pace with the growth in genome sequence data. Metagenomic data is relatively new (see 'The way forward with marine microbial ecology is through an integrated 'meta' approach') and, as a result, functional 'omic' studies of marine microorganisms have almost exclusively been linked to genome sequences of individual organisms.

Mass spectrometry (MS)-based proteomics provides a powerful means of determining proteins expressed under 1 growth condition (proteome snap-shot), or by comparing at least 2 growth conditions (differential expression). Proteomic coverage can be maximized by reducing the complexity of samples through the use of fractionation regimes (e.g. to identify less-abundant proteins) and the sampling of sub-proteomes (e.g. intracellular, membrane, secreted). This type of approach was used to analyse the proteome of the marine bacterium *Alcanivorax borkumensis* in order to determine the metabolic functions involved in petroleum degradation (Sabirova et al. 2006). Whole-cell, soluble fractions are typically used for proteomic analysis. However, it is important to pay attention to other fractions. Secreted proteins may play important roles in antimicrobial activity and cell-cell signalling (Milton 2006). Membrane sub-fractions are technically challenging to analyse but are likely to provide important insight into the mechanisms of how cells sense and respond to their environment. Although 2-dimensional gel electrophoresis (2DE)-based methods are useful (e.g. for fractionation, visualizing post-translational modifications), recent developments in more rapid 'shotgun' approaches using liquid chromatography mass spectrometry (LC-MS) have proven particularly valuable for analysing less-soluble sub-proteomes (e.g. hydrophobic proteins) by providing rapid, high-throughput and broad coverage of these proteins (Wu & Yates 2003, Martosella et al. 2006).

It is not only important to successfully identify proteins, but to accurately quantify protein levels in order to determine the abundance of individual proteins relative to other proteins in the cell (i.e. differential expression). There are 3 major MS-based approaches for quantifying protein levels: 2DE-MS, intensity-based quantification, and stable isotope labeling. Stable isotope labeling is the most comprehensive approach for globally measuring protein abundances and can be performed by *in vitro* (e.g. isotope coded affinity tag: ICAT) and *in vivo* (e.g. metabolic labeling) approaches (Gygi et al. 1999, Krijgsveld et al. 2003, Zhong et al. 2004).

The method developments and refinements of approaches in the field of proteomics (Wilkins et al. 2006) should help to make this technology accessible and immensely useful to the microbial marine biology/ecology field. Reflective of the way in which proteomic methodology has developed, studies of marine microorganisms have primarily involved 2DE-MS approaches (Goodchild et al. 2004a, Gade et al. 2005, Kan et al. 2005, Kim et al. 2005, Sabirova et al. 2006). However, 2DE-MS, LC-MS and ICAT have been applied to *Methanococcoides burtonii* (Goodchild et al. 2004a,b, 2005), and metabolic labeling combined with DNA microarrays with *Methanococcus maripaludis* (Xia et al. 2006). The ability to successfully apply these methods to fastidious, strict anaerobes highlights the potential ease of application of these types of methods to many microorganisms.

Studies of the marine, surface associated bacterium *Pseudoalteromonas tunicata* provides a good example of how genomics and functional genomics can be used to generate new hypotheses about marine ecology. The genome sequence not only provided detailed knowledge of the ability of *P. tunicata* to associate with eucaryal hosts and synthesise novel bioactive metabo-

lites, but enabled comparative proteomic and transcriptomic studies between a wildtype and a mutant (Stelzer et al. 2006). These studies showed that the mutant, which no longer produced bioactives, exhibited an unexpected overexpression of genes involved in iron scavenging and sensing (Stelzer et al. 2006). As a result of these findings, awareness was created about the likely ecological relevance of iron, and this has prompted a series of experimental programs that target the role of iron in bacterial-host interactions in the marine environment.

THE WAY FORWARD WITH MARINE MICROBIAL ECOLOGY IS THROUGH AN INTEGRATED 'META' APPROACH

Tools for characterising the diversity of marine microorganisms have progressed through 3 phases. Initially, marine microbial populations were characterised by isolation and culturing of strains. However, it is now well-accepted that cultivation introduces large qualitative and quantitative biases into ecological studies (Suzuki et al. 1997, Eilers et al. 2000) and only canvases a very small proportion of the total marine diversity (Rappe & Giovannoni 2003). This is primarily a result of the fact that most microorganisms are unable to be cultured using current methods, highlighting the need for access to culture-independent technologies. A second phase arose through molecular ecology studies, which allowed non-culturable bacteria and Archaea to be characterised via molecular fingerprinting, specific molecular probes and sequencing of selected genes (e.g. 16S rDNA and selected functional genes). These PCR-based methods suffer from biased amplification of target sequences and often fail to correctly reflect community composition (Suzuki et al. 1997, Marchesi et al. 1998, von Wintzingerode et al. 1999, Schmalenberger et al. 2001). A third approach, 'metagenomics' (also termed 'environmental genomics', 'ecogenomics' or 'community genomics') has recently emerged, which involves the extraction of DNA from all microorganisms (or size-fractioned components of all microorganisms) from the environment (Handelsmann 2004, Riesenfeld et al. 2004). The eDNA is either cloned as small or large fragments into *Escherichia coli* plasmids and sequenced by the method of Sanger (1977) (e.g. using an ABI 3730 sequencer) (Tyson et al. 2004, Venter et al. 2004), or sequenced directly by high-throughput pyrosequencing (e.g. using a GS20/454 sequencer) (Margulies et al. 2005, Edwards et al. 2006). The eDNA plasmid libraries represent the genomes of the environmental population of microorganisms irrespective of whether the microorganisms are culturable, and can also be used for functional or phylogenetic screening (Handelsmann 2004, Riesenfeld et al. 2004).

The extent of DNA sequencing that is required for the successful reconstruction of genome sequences of microbial communities is directly proportional to the complexity of the environment. Preliminary molecular ecology studies can provide a useful indication of species richness and therefore an estimation of the number of sequencing reactions required. Metagenomic studies of less complex communities are not only less expensive (per sample site) and more easily analysed (e.g. reconstruction of genome sequences of individual species), but are more amenable to metafunctional studies. The metagenome (Tyson et al. 2004) and metaproteome (Ram et al. 2005) study of a biofilm from the acid mine drainage of Iron Mountain is a powerful illustration of what can be achieved in microbial ecology when using a meta-approach. From less than 100 Mb of sequence data, genome sequences for the dominant bacterium (*Leptospirillum* Group II) and archaeon (*Ferroplasma* Type II) and partial genome sequences of several others were obtained. Reconstruction of metabolic pathways led to inferences about nitrogen fixation, which subsequently enabled a successful cultivation strategy to be derived for *Leptospirillum ferrodiazotrophum,* a previously uncultured organism (Tyson et al. 2005).

Metaproteome analysis of the biofilms from the Iron Mountain site led to up to 48% of the predicted proteins being identified from an individual member of the biofilm, a percentage that exceeds the number of proteins typically detected from proteomic studies of microbial isolates. For example, in proteomic studies of Archaea, proteome coverage has been reported as approximately 50% for *Methanocaldococcus jannaschii*, 25% for *Methanococcoides burtonii*, 10% for *Methanosarcina acetivorans* and 34% for *Halobacterium salinarum* (Cavicchioli et al. 2006). In proteomic studies of isolates, monocultures are typically grown in nutrient excess under controlled conditions of growth phase and abiotic influence (e.g. temperature, pH). It is likely that the acid mine biofilm contained cells exhibiting a broad range of phenotypes in response to varying levels of nutrient limitation, interactions with other microorganisms, growth state (e.g. actively growing, dead, planktonic, sessile) and other natural, undefined environmental effectors.

Technological advancement in genome sequencing and proteomics shows no sign of plateauing. Therefore, these meta-approaches will become increasingly feasible for application to more complex environmental samples. Moreover, genomic/proteomic and metagenomic/metaproteomic programs can be run in parallel to more effectively annotate genome sequences and obtain a direct measure of functional gene expression in terms of the presence, relative

abundance and modification states of proteins. The potential of, and challenges for, meta-based studies of microbial communities have been evaluated (Banfield et al. 2005, Foerstner et al. 2006, Ward 2006, Wilmes & Bond 2006), and methods for the preparation of DNA and proteins from soil/sediment or water from environmental samples have been reported (Tsai & Olsen 1991, Purdy et al. 1996, Miller et al. 1999, Schulze 2004, Daniel 2005, Kan et al. 2005, Schulze et al. 2005, Tringe & Rubin 2005).

Isolating representative DNA from communities in soil/sediment is probably one of the most challenging of all natural environmental samples, owing to the complexity and the diversity of microbial populations ($\geq 2 \times 10^4$ bacterial or archaeal species per gram of soil; Daniel 2005), and the fact that microbial cells and free DNA from dead cells adhere to the soil/sediment matrix. DNA can be extracted directly (cells are lysed within sample material) or indirectly (cells are first separated from the environmental sample) (Miller et al. 1999, Daniel 2005), and similar approaches have been adopted for extracting proteins (e.g. Schulze et al. 2005). Marine water samples are easier to manipulate than sediment and tend to have lower complexity. Microbial biomass can be collected by size-fractionated filtration on membrane filters and tangential flow centrifugation, and methods have been developed for subsequent protein extraction and analysis (Schulze 2004, Venter et al. 2004, Kan et al. 2005).

In addition to the use of a meta-approach for obtaining information about microorganisms that are difficult to study (e.g. non-culturable) (Rodriguez-Valera 2004, Tringe & Rubin 2005), proteogenomic studies of individual isolates can also form the basis for rationalizing the need for meta-studies. *Sphingopyxis alaskensis* was isolated as a numerically abundant microorganism from Resurrection Bay in Alaska and oligotrophic waters near Japan, and has served as a model ultramicrobacterium (Cavicchioli et al. 2003). A broad range of laboratory studies including proteomics (e.g. Ostrowski et al. 2004) have defined physiological characteristics that distinguish it from typical copiotrophic bacteria, such as *Photobacterium angustum* S14 (Cavicchioli et al. 2003). Despite being isolated by extinction dilution methods and representing a numerically abundant organism at the time of sampling, *S. alaskensis* has not been reported to be as widely distributed as SAR11, which is apparently one of the most cosmopolitan microorganisms in oligotrophic oceanic waters. Metagenomics of distinct oceanic sites along the path of the Sorcerer II expedition (Venter et al. 2004) have revealed an astounding level of total microbial genetic diversity. To date, metagenome studies have not included North Pacific waters where *S. alaskensis* was isolated. It has been proposed that *S. alaskensis* may circulate between locations that are 10 000 km apart by ocean currents in the North Pacific (Eguchi et al. 2001), and it will be valuable to assess the genomic variation that exists between populations of *S. alaskensis* from the geographically distinct regions of the North Pacific from where it was previously isolated. Moreover, when the analysis of *S. alaskensis* genome sequence is complete, similarities and differences with SAR11 will be able to be documented. It is already clear that *S. alaskensis* has a significantly larger genome (~3.2 Mb) than SAR11 (1.3 Mb). It has previously been argued that multiple strategies may have evolved to enable microorganisms to compete effectively in oligotrophic waters (Cavicchioli et al. 2003). It will be valuable to assess the metaproteome of both *S. alaskensis* and SAR11 in their native environments to determine which component of their genetic complement is expressed, and to infer how this may affect their individual adaptation strategies.

The value of individual microorganisms guiding metagenome studies is also well illustrated by studies of psychrophilic Archaea. Cold-adapted Archaea perform diverse functional roles in a wide range of cold environments, and the extent to which they transform the cold biosphere can be appreciated from their phylogenetic and functional diversity, abundance and range of cold biotopes they inhabit (Cavicchioli 2006). They represent an important fraction of cold marine environments and have been detected in deep ocean waters and sediment, sea ice and marine-derived Antarctic lakes. *Methanococcoides burtonii* was isolated from a marine-derived lake (Ace Lake) in Antarctica, and through studies of cold adaptation that addressed protein structure, intracellular solutes, membrane lipids, tRNA modification, gene regulation, comparative genomics and proteomics, it has developed into the model psychrophilic archaeon (Cavicchioli 2006). In addition to *M. burtonii*, *Methanogenium frigidum* (Ace Lake) and *Halorubrum lacusprofundi* (Deep Lake) were isolated from Antarctica. The studies of these individual isolates from Antarctic lakes have generated a broad range of questions that can be addressed most successfully by performing metagenomic and associated functional studies. The types of questions include: are genes that are preferentially expressed at 4°C under laboratory growth conditions (and therefore thought to be involved in cold adaptation) expressed in the environment (Goodchild et al. 2004a, 2005)? Are genes that have been linked to genome plasticity (e.g. transposons) (Goodchild et al. 2004b) expressed in the environment, and how does this affect the overall microheterogeneity of species such as *M. burtonii*, *M. frigidum* and *H. lacusprofundi*? Which hypothetical proteins are synthesized and

therefore important for growth of the organism in the environment (Saunders et al. 2005)?

By coupling metagenomic analysis with a range of other experiments performed at the time of sampling, including isolations, physical and chemical measurements and labeling experiments, and integrating this with established physical and geochemical data of the lakes, a comprehensive understanding of the microbial system can be defined. The metagenomics will particularly facilitate the ability to define community structure, individuals within the communities and their associated biological processes. Key biological properties that underpin the biogeochemical process will also be able to be derived from the genomic properties of key microbial groups (e.g. methanogens and methylotrophs), as will an understanding of how key chemical cycles (e.g. methane) are microbially driven, and of what biological properties define life at abiotic limits (e.g. cold, hypersalinity, oligotrophy). While these examples pertain to specific Antarctic lakes and the archaeal isolates, the principles of the approaches are applicable to other environments and individual isolates from those environments. This clearly illustrates how studies of cultivated organisms can be greatly facilitated by subsequent metagenomic studies, in addition to the obvious advantages metagenomics offers to studies of uncultivated species in their respective environments.

PERSPECTIVE

Metagenomic studies have identified the high level of microheterogeneity that can exist within populations. One of the first studies of this type documented the existence of 2 major variants of *Cenarchaeum symbiosum* that coexist in a marine sponge (Schleper et al. 1998). However, it is not clear what biological and abiotic factors control the extent and tempo of genomic heterogeneity. Metagenomics of samples from the Sargasso Sea highlighted the microheterogeneity within marine *Prochlorococcus marinus* populations (Venter et al. 2004), even though it is not clear over what time period this has occurred. Biofilms can exhibit large changes in genetic composition over their life-time (e.g. within a few days) (Webb et al. 2004, Mai-Prochnow et al. 2006), and may be major ecological drivers of genetic diversity in 'real-time', and hence model systems for studying genome evolution. In fact, natural biofilm populations have been shown to not contain discrete genome sequences for a particular species, but rather to possess a highly diverse, mosaic genomic complement (Tyson et al. 2004, Allen & Banfield 2005). In contrast to this dynamic system, microheterogenous communities that double at very slow rates (e.g. deep subsurface) may be considered repositories of genetic codes, rather than as genomic solutions that have evolved subject to the influences of modern-day pertubations.

Phylogenetic analysis of 16S rRNA genes was largely responsible for the revolution in evolutionary biology that led to the definition of the 3 domains of life (Woese et al. 1990), and to the great expansion of the diversity of known species. Genome sequencing of individual microorganisms corroborated the concept of the 3 domains of life, and uncovered the extent to which genetic diversity exists within apparently coherent species (e.g. *E. coli*). Metagenomics is revealing the extent to which genetic diversity exists within natural communities, and is challenging the concept of a microbial species. By understanding the extent, tempo and mode of genome evolution it will be possible to gain a practical understanding of microbial community evolution and infer the effects of human impact on the microbial gene pool, and greatly enrich our understanding of how life has evolved along its 3.8 billion yr old trajectory.

LITERATURE CITED

Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3:489–498

Ando S, Ishikawa K, Ishida H, Kawarabayasi Y, Kikuchi H, Kosugi Y (1999) Thermostable aminopeptidase from *Pyrococcus horikoshii*. FEBS Lett 447:25–28

Banfield JF, Verberkmoes NC, Hettich RL, Thelen MP (2005) Proteogenomic approaches for the molecular characterization of natural microbial communities. OMICS 9: 301–333

Beja O, Aravind L, Koonin EV, Suzuki MT and 8 others (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906

Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. Nature 411:786–789

Cavicchioli R (2006) Cold adapted Archaea. Nature Rev Microbiol 4:331–343

Cavicchioli R, Ostrowski M, Fegatella F, Goodchild A, Guixa-Boixereu N (2003) Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of *Sphingopyxis alaskensis* (formerly *Sphingomonas alaskensis*). Microb Ecol 45:203–217

Cavicchioli R, Goodchild A, Raftery M (2006) Proteomics of Archaea (Chapter 5). In: Humphery-Smith I, Hecker M (eds) Microbial proteomics — functional biology of whole organisms. Wiley, New York

Daniel R (2005) The metagenomics of soil. Nature Rev Microbiol 3:470–478

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M and 6 others (2006) Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. BMC Genomics 7:57

Eguchi M, Ostrowski M, Fegatella F, Bowman J, Nichols D, Nishino T, Cavicchioli R (2001) *Sphingomonas alaskensis*, strain AF01: an abundant oligotrophic ultramicrobacterium from the North Pacific. Appl Environ Microbiol 67: 4945–4954

Eilers H, Pernthaler J, Gloeckner FO, Amann R (2000) Culturability and *in situ* abundance of pleagic bacteria form the North Sea. Appl Environ Microbiol 66:3044–3051

Foerstner KU, von Mering C, Bork P (2006) Comparative analysis of environmental sequences: potential and challenges. Phil Trans R Soc Lond B 361:519–523

Gade D, Theiss D, Lange D, Mirgorodskaya E and 8 others (2005) Towards the proteome of the marine bacterium *Rhodopirellula baltica*: Mapping the soluble proteins. Proteomics 5:3654–3671

Giovannoni SJ, Tripp HJ, Givan S, Podar M and 10 others (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245

Goldberg SM, Johnson J, Busam D, Feldblyum T and 15 others (2006) A sanger/pyrosequencing hybrid approach for generation of high quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA 103: 11240–11245

Goodchild A, Saunders NFW, Ertan H, Raftery M, Guilhaus M, Curmi PMG, Cavicchioli R (2004a) A proteomic determination of cold adaptation in the Antarctic archaeon, *Methanococcoides burtonii*. Mol Microbiol 53:309–321

Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2004b) The biology of the cold adapted archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. J Proteome Res 3:1164–1176

Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2005) Cold adaptation of the Antarctic archaeon, *Methanococcoides burtonii* assessed by proteomics using ICAT. J Proteome Res 4:473–480

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnol 17: 994–999

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM, DeLong EF (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. PLoS Biol 4:e95

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669–685

Kan J, Hanson TE, Ginter JM, Wang K, Chen F (2005) Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Syst 19:7–19

Kim YK, Yoo WI, Lee SH, Lee MY (2005) Proteomic analysis of cadmium-induced protein profile alterations from marine alga *Nannochloropsis oculata*. Ecotoxicol 14:589–596

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543–546

Krijgsveld J, Ketting RF, Mahmoudi T, Johansen J, Artal-Sanz M, Verrijzer CP, Plasterk RHA, Heck AJR (2003) Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. Nature Biotechnol 21:927–931

Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34: D332–334

Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glockner FO (2006) Megx.net—database resources for marine ecological genomics. Nucleic Acids Res 34:D390–393

Mai-Prochnow A, Webb JS, Ferrari BC, Kjelleberg S (2006) Ecological advantages of autolysis during biofilm development and dispersal of *Pseudoalteromonas tunicata.* Appl Environ Microbiol 72:5414–5420

Marchesi JR, Sato T, Weightman AJ, Martin TA, Fry JC, Hiom SJ, Dymock D, Wade WG (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. Appl Environ Microbiol 64: 795–799

Margulies M, Egholm M, Altman WE, Attiya S and 52 others (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Martosella J, Zolotarjova N, Liu H, Moyer SC, Perkins PD, Boyes BE (2006) High recovery HPLC separation of lipid rafts for membrane proteome analysis. J Proteome Res 5: 1305–1312

Miller DN, Bryant JE, Madsen EL, Ghiorse WC (1999) Evaluation and optimization of DNA extraction and purification procedures of soil and sediment samples. Appl Environ Microbiol 65:4715–4724

Milton DL (2006) Quorum sensing in vibrios: complexity for diversification. Int J Med Microbiol 296:61–71

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF and 31 others (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432:910–913

Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 21:4322–4329

Ostrowski M, Fegatella F, Wasinger V, Corthals G, Guilhaus M, Cavicchioli R (2004) Cross species identification of proteins from proteome profiles of the marine oligotrophic ultramicrobacterium, *Sphingopyxis alaskensis*. Proteomics 4:1779–1788

Purdy KJ, Embley TM, Takii S, Nedwell DB (1996) Rapid extraction of DNA and rRNA from sediments by a novel hydroxyapatite spin-column method. Appl Environ Microbiol 62:3905–3907

Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW and 5 others (2005) Community proteomics of a natural microbial biofilm. Science 308:1915–1920

Rappe M, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394

Riesenfeld C, Schloss P, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525–552

Rodríguez-Valera F (2004) Environmental genomics, the big picture? Microbiol Lett 231:153–158

Sabirova JS, Ferrer M, Regenhardt D, Timmis KN, Golyshin PN (2006) Proteomic insights into metabolic adaptations in *Alcanivorax borkumensis* induced by alkane utilization. J Bacteriol 188:3763–3773

Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21:4320–4321

Sanger F, Nicklen S, Coulsen AR (1977) DNA sequencing with chain termination. Proc Natl Acad Sci USA. 74: 5463–5467

Saunders NFW, Goodchild A, Raftery M, Guilhaus M, Curmi PMG, Cavicchioli R (2005) Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanococcoides burtonii*. J Proteome Res 4:464–472

Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 180:5003–5009

Schleper C, Jurgens G, Jonuscheit M (2005) Genomic studies of uncultivated archaea. Nat Rev Microbiol 3:479–488

Schmalenberger A, Schwieger F, Tebbe CC (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl Environ Microbiol 67:3557–3563

Schulze W (2004) Environmental proteomics—what proteins from soil and surface water can tell us: a perspective. Biogeosciences 1:195–218

Schulze WX, Gleixner G, Kaiser K, Guggenberger G, Mann M, Schulze ED (2005) A proteomic fingerprint of dissolved organic carbon and of soil particles. Oecologia 142: 335–343

Stelzer S, Egan S, Larsen MR, Bartlett DH, Kjelleberg S (2006) Unravelling the role of the ToxR-like transcriptional regulator WmpR in the marine antifouling bacterium *Pseudoalteromonas tunicata*. Microbiology 152:1385–1394

Suzuki MT, Rappe MS, Haimberger ZW, Winfield H, Adair N, Strobel J, Giovannoni SJ (1997) Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. Appl Environ Microbiol 63: 983–989

Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. Nature Rev Microbiol 6: 805–814

Tsai YL, Olson BH (1991) Rapid method for direct extraction of DNA from soil and sediments. Appl Environ Microbiol 57:1070–1074

Tyson GW, Chapman J, Hugenholtz P, Allen EE and 6 others (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF (2005) Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. Appl Environ Microbiol 71:6319–6324

Venter J, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

von Wintzingerode F, Gobel UB, Stackebrandt E (1999) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev 21:213–219

Ward N (2006) New directions and interactions in metagenomics research. FEMS Microbiol Ecol 55:331–338

Webb JS, Lau M, Kjelleberg S (2004) Bacteriophage and phenotypic variation in *Pseudomonas aeruginosa* biofilm development. J Bacteriol 186:8066–8073

Wilkins MR, Appel RD, Van Eyk JE, Chung MCM and 12 others (2006) Guidelines for the next 10 years of proteomics. Proteomics 6:4–8

Wilmes P, Bond PL (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol 14:92–97

Woese CR, Kandler O, Wheelis ML (1990) Toward a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87: 4576–4579

Wu CC, Yates JR (2003) The application of mass spectrometry to membrane proteomics. Nature Biotechnol 21:262–267

Xia Q, Hendrickson EL, Zhang Y, Wang T and 6 others (2006) Quantitative proteomics of the archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR. Mol Cell Proteomics 5:868–881

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. Nature Biotechnol 24:680–686

Zhong H, Marcus SL, Li L (2004) Two-dimensional mass spectra generated from the analysis of ^{15}N-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing. J Prot Res 3:1155–1163

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: June 29, 2006; Accepted: July 28, 2006
Proofs received from author(s): February 7, 2007

Vol. 332: 301–306, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Metabolomics of aquatic organisms: the new 'omics' on the block

Mark R. Viant*

School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

ABSTRACT: Environmental metabolomics can be defined as the application of metabolomics to characterise the metabolism of free-living organisms obtained from the natural environment and of organisms reared under laboratory conditions, where those conditions serve to mimic scenarios encountered in the natural environment. This approach has considerable potential for characterising the responses of organisms to natural and anthropogenic stressors. The current essay introduces environmental metabolomics, discusses the challenges of measuring metabolites, and then highlights the dynamic nature of the metabolome that can be exploited to provide a holistic view of an organism's health. Dealing with metabolic variability is a considerable challenge in environmental metabolomics. Here, I propose the concept of a normal metabolic operating range (NMOR), defined as the region in metabolic space in which 95% of individuals from a population reside, with stress identified as a deviation from the NMOR. Furthermore, I emphasise the importance of genotypic and phenotypic anchoring (e.g. knowing species, gender, age) to facilitate interpretation of multivariate metabolomics data.

KEY WORDS: Metabolomics · Environmental metabolomics · Environment · Phenotypic anchoring · Normal metabolic operating range · Fish · Invertebrate · Stress

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

Metabolomics is the study of the endogenous low molecular weight metabolites within a cell, tissue or biofluid (termed the metabolome). 'Environmental metabolomics' is the application of metabolomics to characterise the metabolic responses of an organism to both natural and anthropogenic stressors that can occur in its environment. Although it is the newest 'omic' approach, and is therefore considerably less developed and utilised in marine environmental research than genomics, transcriptomics and proteomics, it potentially affords several benefits for assessing organism function and health at the molecular level. For example, metabolomics shares many of the advantages of the other omics approaches in that it enables a rapid, unbiased and simultaneous measurement of many tens, hundreds or even a thousand endpoints (i.e. metabolites), and therefore differs substantially from traditional biochemical methods that typically detect only 1 or a few metabolites. As a result, metabolomics is a particularly powerful approach for discovering biomarker profiles of toxicant exposure and disease, and for identifying the metabolic pathways involved in such processes. Metabolomics now offers us a systems-based approach for studying individuals in the marine environment. Other advantages that are specific to metabolomics include the high degree of functionality of metabolic measurements that can be directly related to an organism's phenotype, and the flexibility with which it can be applied to any organism irrespective of the knowledge of the genome for that species.

To date there have been only 10 publications that have applied metabolomics to aquatic species, which can be grouped into the study of biological stress (i.e. disease), chemical stress (i.e. toxicity), temperature stress (Viant et al. 2003a) and fish embryogenesis (Viant 2003). The toxicity studies include exposure of embryos of medaka *Oryzias latipes* (Viant et al. 2005, Viant et al. 2006a) and chinook salmon *Oncorhynchus tshawytscha* (Viant et al. 2006b) with the goal of evalu-

*Email: m.viant@bham.ac.uk

© Inter-Research 2007 · www.int-res.com

ating metabolomics as a high-throughput screening tool for chemical risk assessment. Samuelsson et al. (2006) have utilised metabolomics to study the effects of ethinylestradiol, an endocrine disruptor, in juvenile rainbow trout *Oncorhynchus mykiss*. The effects of diseases on the metabolome have been studied in fish and an invertebrate, including a bacterial infection in Atlantic salmon *Salmo salar* (Solanky et al. 2005), liver cancer in dab *Limanda limanda* (Stentiford et al. 2005), and withering syndrome in red abalone *Haliotis rufescens* (Viant et al. 2003b, Rosenblum et al. 2005). Unexplored applications of metabolomics that could benefit from this rapid, unbiased systems-based approach include its use in environmental monitoring and in the aquaculture industry to optimise husbandry and productivity.

An important characteristic that in many situations sets the metabolome apart from the genome, transcriptome and proteome is the degree to which it varies under normal and stressful conditions. The metabolome is often the first to respond to anthropogenic stressors (e.g. pollutant exposure) and natural daily events (e.g. feeding), and in some cases no changes in the transcriptome and proteome occur. Furthermore, these metabolic changes can occur directly (e.g. oxidative stress associated with antioxidant depletion) or indirectly (e.g. by redistribution of energy reserves away from growth and reproduction towards cellular defence and repair). As discussed later, this large variation in the metabolome has a number of major ramifications for the applicability of metabolomics in environmental studies. First, however, it is important to address and clarify some issues relating to the measurement of metabolites.

MEASUREMENT OF THE METABOLOME

As stated above, metabolomics is the newest of the omic approaches and is still very much under development. This is particularly true for the methods used to measure metabolite levels. Metabolites, unlike genes, transcripts and proteins, are a highly physically and chemically diverse group of chemicals. Some, like glycine, are present at high concentration, have a low molecular mass and are extremely polar. Others, such as testosterone, have the opposite characteristics. Several thousand additional metabolites have intermediate or even more extreme properties. So, unlike the measurement of genes, transcripts and to some extent proteins (which are all polymers of nucleotide bases or amino acids), no one bioanalytical technique is capable of detecting all metabolites.

The 2 most widely used methods in metabolomics are ^{1}H nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry. A comparison of these techniques is beyond the scope of this commentary, so readers are referred to articles by Dunn & Ellis (2005), Pelczer (2005), and Villas-Boas et al. (2005). It is important to note, however, that both techniques have considerable value in metabolomics and that neither method has yet been fully developed for this application. Perhaps the most striking statistic that illustrates this point is that of the estimated several thousand metabolites in the cellular metabolome; current NMR methods are believed to detect only about 100 metabolites (less than 10%) and mass spectrometry up to approximately 1000 metabolites. It is therefore important for those engaged in environmental metabolomics research to stay acquainted with technological advances in this rapidly developing field, and to implement them as they occur. The 2 areas that are in most urgent need of development include methods to extend the coverage of the metabolome and an improved ability to identify and quantify metabolites unambiguously. Before leaving this topic, it is important to recognise a definitive advantage associated with measuring metabolites that stems from the conservation of metabolites across species. Any technical advances in the measurement of metabolites in one species will in general be applicable to all other species, and no *a priori* genomic knowledge is required. Exploitation of this fact is discussed below.

THE DYNAMIC METABOLOME

One of the most significant advantages of metabolomics is the close dynamic relationship that exists between the metabolites that are measured and the physiological status of the whole organism. For example, metabolomics includes the measurement of ATP and glycogen, which can vary as a function of the energetic status of an organism (Wasser et al. 1996). Metabolomic methods can also detect molecules like glutathione and ascorbate, which can change as a function of cellular redox status (Kristal et al. 1998). Steroids are another class of molecules that are of considerable interest, and so measurement of oestradiol and ketotestosterone could be used to help inform on the reproductive status of an organism (Noaksson et al. 2004). Taken together, as metabolomic technologies are developed to the point where many hundreds of metabolites are measured simultaneously, the exciting potential to rapidly assess many aspects of an organism's current energetic, oxidative and perhaps even reproductive status may be realised. Measurement of the genome and transcriptome are less able to provide this information, because genes and transcripts are not guaranteed to manifest themselves as functional

changes at the organismal level. Conversely, metabolomics is not useful for assessing population structure and genealogy of marine organisms, for which genomics is vital. Although the proteome can provide a window into functional organismal changes, the measurement of protein levels is still not able to provide such a direct link to physiology (such as energetic status) as can be achieved via the metabolome. The attempt to prove a causal relationship between metabolic biomarker profiles and an individual's Darwinian fitness (reproductive health, growth and survival) is, therefore, an important area of current and future research.

The ability of the metabolome to change so readily is not only a considerable strength but also creates a major challenge. Measurement of the metabolomes of several individuals from a free-living population will necessarily include considerable metabolic 'noise', i.e. the metabolite concentrations will be highly variable among individuals owing to differences in the individual's local environment, their genetics, and possibly the time since they last ate! This biological noise will tend to mask the metabolic differences between healthy and stressed animals, as well as the more subtle differences among closely related stressful states. Coping with this biological noise is in my view the greatest challenge in environmental metabolomics. There are, fortunately, a number of approaches that will help address this problem, including the simultaneous measurement of multiple metabolites, supervised methods of multivariate analysis, and knowledge of the organism under investigation, which are all addressed below.

MEASUREMENT OF MULTIPLE METABOLITES, NORMAL OPERATING RANGES AND MULTIVARIATE MODELS

Biological variability has long been the thorn in the side of the environmental biomarker research community. Numerous studies have reported that seemingly well established biomarkers for pollutant exposure such as metallothionein, heat shock proteins and antioxidant defence mechanisms exhibit seasonal variability (Sheehan & Power 1999, Geffard et al. 2001, Lacorn et al. 2001). This has limited the application of biomarkers as a tool for ecological monitoring. Metabolomics, as with all the omics, brings an interesting new dimension to biomarker research, that being the simultaneous measurement of potentially 100s or 1000s of metabolites. It might be construed that this exacerbates the problem because the total amount of variability captured in these multiple measurements will be significantly greater than the variability in any one metabolite. However, the advantage of the simultaneous measurement of multiple parameters stems from the fact that there will be a subset of metabolites within all those measured that can each (partially) discriminate between healthy and stressed organisms. The integrated profile of this subset of metabolic biomarkers will be able to discriminate between the healthy and stressed groups more robustly than any one biomarker alone. In effect, the biomarker profile becomes stabilised by the inclusion of many relevant variables, even if each of these variables is noisy (Eriksson et al. 2001). The challenge is to determine which subset of the hundreds of metabolites is able to provide this discrimination, which I address below.

Aside from the advantage of providing subsets of biomarkers that are potentially specific to a defined stressor, the simultaneous measurement of multiple metabolites enables us to obtain a more holistic view of the metabolic status of individuals in a population. That is, multiple measurements allow us to determine the 'normal operating range' (NOR) of an aquatic organism, a concept discussed by Kersting (1984). The metabolic status of an organism is necessarily a multivariate property in which the concentrations of each of the few thousand metabolites that define the metabolome are represented along unique axes in multidimensional metabolic space. The normal metabolic operating range (NMOR) can be defined as the region in that space in which 95% of the individuals from a population reside. Stress can then be defined as a deviation from the NMOR, with different stressors inducing different metabolic responses and therefore moving away from the NMOR in unique directions. Since we are unable to visualise high dimensional space, we use dimensionality reduction tools such as principal components analysis (PCA; Eriksson et al. 2001) to project multi-dimensional space down to just a few dimensions. This is illustrated in Fig. 1, which shows the PCA scores plot for a cohort of 15 marine mussels *Mytilus galloprovincialis* from Port Quin, Cornwall, UK. These mussels had been submerged for at least 2 h, were actively respiring, and were then dissected and the adductor muscles rapidly frozen. Metabolites were extracted from the muscle, analysed by NMR spectroscopy, and the resulting spectra were subject to PCA. In the PCA scores plot (Fig. 1), the ellipse — drawn at 2 SD from the mean metabolic status of the adductor muscles — represents the NMOR. The NMOR concept provides a useful and visibly meaningful approach for summarising high dimensional metabolomics data.

In conjunction with multivariate analyses, the concept of NMOR can be extended to help visualise the effect of stressors. When comparing stressed and unstressed animals, some of the metabolites with the

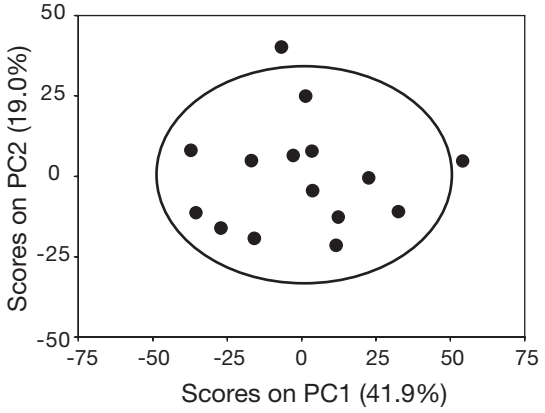

Fig. 1. *Mytilus galloprovincialis*. PCA scores plot from analysis of the metabolic fingerprints of adductor muscle from 15 mussels collected from Port Quin, Cornwall. Each data point corresponds to an entire NMR metabolic fingerprint comprising ca. 100 metabolites, and PCA axes are linear combinations of the most variable metabolites. If 2 data points are closely spaced, then this indicates that metabolomes of those samples are similar. The ellipse (±2 SD) defines the normal metabolic operating range (NMOR) for 95% of individuals within the population; % variance accounted for by each PC is shown

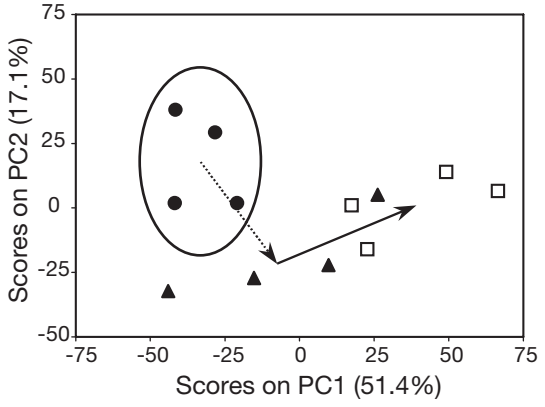

Fig. 2. *Mytilus galloprovincialis*. PCA scores plot from analysis of the NMR metabolic fingerprints of adductor muscle from mussels exposed to 0 (●), 250 (▲) and 1000 (□) ppb copper for 48 h. The NMOR of the control group is shown by the ellipse (±2 SD). Copper-induced stress forces the average metabolic status of the mussels away from the NMOR, as indicated by the dotted (low dose) and solid (high dose) arrow

most variable concentrations will be completely unrelated to the stress, and these will tend to mask those induced by the stressor. The application of unsupervised methods of analysis[1] such as PCA will only identify the most variable metabolites, irrespective of whether they are related to the stressor. Occasionally, when the induced stress is large relative to the biological noise, PCA will detect the differences between the metabolic phenotypes of the 2 groups. An example is given in Fig. 2, which shows the effect of a 48 h exposure of mussels to 0, 250 and 1000 ppb copper. Despite the small number of samples in this pilot study, a dose-dependent deviation from the NMOR is clearly visible. For the more typical scenario in which the induced stress is small relative to the biological noise, a statistically more powerful approach is required for determining specific biomarkers of stress. Supervised methods of analysis[2] can be used to search specifically for those metabolites that discriminate the stressed and unstressed groups (assuming sample sizes are sufficiently large). It is highly likely that supervised methods will be essential for characterising the effects of stressors in free-living aquatic organisms. Supervised methods for robust classification of metabolic phenotype and for biomarker discovery include partial least-squares discriminant analysis (Eriksson et al. 2001) and genetic algorithms (Jarvis & Goodacre 2005). Results from these analyses must be accompanied by appropriate parameters such as cross-validation misclassification rates, sensitivity and specificity, which can assess the quality of the multivariate model. This and the earlier discussions highlight a considerable challenge in environmental metabolomics, namely that knowledge in several disciplines spanning the ecology and biology of the organism through to sophisticated bioanalytical and bioinformatic techniques must be mastered, which necessitates collaboration between research groups. Furthermore, owing to the expense of establishing, and expertise required to operate, a metabolomics bioanalytical laboratory, a logical strategy is to establish centralised facilities that act as centres of excellence in environmental metabolomics.

IMPORTANCE OF GENOTYPIC AND PHENOTYPIC ANCHORING

From my laboratory's studies on field-sampled mussels it has become evident that, in order to elucidate metabolic biomarker profiles for specific stressors, we need a complete phenotypic and genotypic characterisation of these animals. For example, metabolomics studies of rodents established that urinary metabolite composition depends upon strain (Gavaghan et al. 2000), sex (Stanley et al. 2005) and age (Plumb et al. 2003). It is logical to conclude that similar genotypic

[1]Unsupervised analyses do not use class identifiers (e.g. control or diseased). They aim to detect clusters in the metabolic data that may not be trivially observable and that indicate which animals have similar metabolomes

[2]Supervised analyses do use class identifiers. The aim is to build a multivariate model that can predict those classifications (e.g. can discriminate between healthy and diseased animals) and discover relevant biomarkers

and phenotypic traits are important for understanding changes in the metabolome of mussels. To date, however, toxicity studies with field-sampled mussels from the UK have mostly been conducted without regard to these traits. This is particularly worrying when one considers that the UK coastline is populated by the native *Mytilus edulis* and the Mediterranean *M. galloprovincialis*, as well as a viable hybrid species (Hilbish et al. 2002). As such, we first need to assess the effects of species, sex and age on the metabolome, prior to characterising the metabolic responses of these animals to environmental stressors. That is, we need to deconvolute the overall biological noise into components with regard to major phenotypic and genotypic traits, so that these can effectively be eliminated. Of course we will never be able to characterise all noise, but to understand (and to effectively anchor to known traits) just some of it will lessen the computational challenge associated with finding biomarkers to specific stressors. Furthermore, characterising the metabolic effect of natural stressors such as hypoxia and food limitation will also be important for helping to unravel the effects of natural stressors, anthropogenic stressors and residual biological noise. Since this is true for all environmental metabolomics studies, I recommend that as metabolomics becomes more widely used in marine ecology, the baseline biochemistry of aquatic animals should be much more thoroughly characterised as a function of species, sex, age, reproductive cycle and the effects of natural stressors. For many species this will require studying the metabolic changes throughout an entire annual cycle. Such studies have the potential to add significantly to our knowledge of these organisms and their interactions with the environment.

An additional complication arises when comparing the metabolomes of organisms collected from multiple sites. It is quite plausible that comparison among these organisms will show metabolic differences that arise from, for example, differences in food availability or differences in temperature that affect the timing of the reproductive cycle. This will complicate the interpretation of the metabolomics data because these differences may mask effects resulting from anthropogenic stressors. One approach to lessen this problem builds on the concept of phenotypic anchoring, and that is to additionally anchor the metabolic measurements with multiple physical and chemical descriptors of each site. A more robust solution would be to conduct temporal studies at a series of independent sites, and to use each of those sites as its own internal control. This could then enable temporal changes in environmental quality at these sites to be assessed via changes in the metabolomes of the resident sentinel species.

TOWARDS MULTI-SPECIES ASSESSMENT OF ECOLOGICAL HEALTH

The conservation of metabolites among species (i.e. alanine is conserved in marine mammals, fish and invertebrates) provides metabolomics with a significant advantage for multi-species assessment compared with the other omics that rely on species-specific information. Metabolomics could therefore assess the metabolic status and derive NMORs for multiple species at one geographical location. Using the same argument as above—that the measurement of multiple metabolites (versus one) can provide a more robust assessment of organismal metabolic health—one can argue that characterising the health of multiple species can provide a more complete assessment of ecosystem responses to environmental stressors. This approach was previously used to analyse invertebrate community responses to an anionic surfactant in stream mesocosms, but in that case the invertebrate populations were simply counted (Wong et al. 2003). Metabolomics could afford a more sensitive window into the deterioration of organism health (prior to death) and could be used to identify which species in the community are most sensitive. Furthermore, the specific metabolic changes observed could help to inform on the nature of the stressor.

CONCLUSIONS

Metabolomics, being the new omics on the block, has benefited from some of the lessons learned from the development of the other omics. This is particularly true in terms of the recognition that sophisticated multivariate analyses and data standardisation are both essential, together with the acceptance (by some) of the significant value of unbiased discovery driven research. Unfortunately, metabolomics has also been labelled with the same shortfalls and viewed by some with the same scepticism applied to the other omics. But at this stage is this justified? Considering that only a handful of research laboratories are engaged in aquatic environmental metabolomics, and given the small amount of funding that has been available to date, it is not surprising that there are only 10 publications in the field. While we do not yet know to what extent metabolomics will impact on marine ecological studies, what is definitely true is that the approach offers considerable potential for rapid assessment of the metabolic status of marine organisms, is capable of multi-species investigations, and can provide molecular information that is closely related to whole-organism physiology and function. Ultimately, we should strive towards the integration of omics data sets because this will enable us to

exploit the advantages of each approach and will provide the most comprehensive molecular description of organisms in the aquatic environment.

Acknowledgements. I thank the Natural Environment Research Council, UK, for an Advanced Fellowship in Metabolomics (NER/J/S/2002/00618). I am grateful to A. Hines and S. Oladiran for the mussel data and to Professor K. Chipman for providing feedback on the manuscript.

LITERATURE CITED

Dunn WB, Ellis DI (2005) Metabolomics: current analytical platforms and methodologies. Trends Anal Chem 24: 285–294

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi- and megavariate data analysis—principles and applications. Umetrics, Umea

Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK (2000) An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. FEBS Lett 484:169–174

Geffard A, Amiard-Triquet C, Amiard JC, Mouneyrac C (2001) Temporal variations of metallothionein and metal concentrations in the digestive gland of oysters (*Crassostrea gigas*) from a clean and a metal-rich site. Biomarkers 6:91–107

Hilbish TJ, Carson EW, Plante JR, Weaver LA, Gilg MR (2002) Distribution of *Mytilus edulis*, *M. galloprovincialis*, and their hybrids in open-coast populations of mussels in southwestern England. Mar Biol 140:137–142

Jarvis RM, Goodacre R (2005) Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. Bioinformatics 21:860–868

Kersting K (1984) Normalized ecosystem strain—a system parameter for the analysis of toxic stress in (micro) ecosystems. Ecol Bull 36:150–153

Kristal BS, Vigneau-Callahan KE, Matson WR (1998) Simultaneous analysis of the majority of low-molecular-weight, redox-active compounds from mitochondria. Anal Biochem 263:18–25

Lacorn M, Piechotta G, Simat TJ, Kammann U and 6 others (2001) Annual cycles of apoptosis, DNA strand breaks, heat shock proteins, and metallothionein isoforms in dab (*Limanda limanda*): influences of natural factors and consequences for biological effect monitoring. Biomarkers 6: 108–126

Noaksson E, Gustavsson B, Linderoth M, Zebuhr Y, Broman D, Balk L (2004) Gonad development and plasma steroid profiles by HRGC/HRMS during one reproductive cycle in reference and leachate-exposed female perch (*Perca fluviatilis*). Toxicol Appl Pharmacol 195:247–261

Pelczer I (2005) High-resolution NMR for metabomics. Curr Opin Drug Discov Dev 8:127–133

Plumb R, Granger J, Stumpf C, Wilson ID, Evans JA, Lenz EM (2003) Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. Analyst 128:819–823

Rosenblum ES, Viant MR, Braid BM, Moore JD, Friedman CS, Tjeerdema RS (2005) Characterizing the metabolic actions of natural stresses in the California red abalone, *Haliotis rufescens* using ^{1}H NMR metabolomics. Metabolomics 1: 199–209

Samuelsson LM, Förlin L, Karlsson G, Adolfsson-Erici M, Larsson DGJ (2006) Using NMR metabolomics to identify responses of an environmental estrogen in blood plasma of fish. Aquat Toxicol 78:341–349

Sheehan D, Power A (1999) Effects of seasonality on xenobiotic and antioxidant defence mechanisms of bivalve molluscs. Comp Biochem Physiol 123C:193–199

Solanky KS, Burton IW, MacKinnon SL, Walter JA, Dacanay A (2005) Metabolic changes in Atlantic salmon exposed to *Aeromonas salmonicida* detected by ^{1}H nuclear magnetic resonance spectroscopy of plasma. Dis Aquat Org 65: 107–114

Stanley EG, Bailey NJC, Bollard ME, Haselden JN, Waterfield CJ, Holmes E, Nicholson JK (2005) Sexual dimorphism in urinary metabolite profiles of Han Wistar rats revealed by nuclear magnetic resonance based metabonomics. Anal Biochem 343:195–202

Stentiford GD, Viant MR, Ward DG, Johnson PJ and 5 others (2005) Liver tumours in wild flatfish: a histopathological, proteomic and metabolomic study. Omics 9:281–299

Viant MR (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Comm 310:943–948

Viant MR, Werner I, Rosenblum ES, Gantner AS, Tjeerdema RS, Johnson ML (2003a) Correlation between heat-shock protein induction and reduced metabolic condition in juvenile steelhead trout (*Oncorhynchus mykiss*) chronically exposed to elevated temperature. Fish Physiol Biochem 29:159–171

Viant MR, Rosenblum ES, Tjeerdema RS (2003b) NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. Environ Sci Technol 37:4982–4989

Viant MR, Bundy JG, Pincetich CA, de Ropp JS, Tjeerdema RS (2005) NMR-derived developmental metabolic trajectories: an approach for visualizing the toxic actions of trichloroethylene during embryogenesis. Metabolomics 1: 149–158

Viant MR, Pincetich CA, Hinton DE, Tjeerdema RS (2006a) Toxic actions of dinoseb in medaka (*Oryzias latipes*) embryos as determined by *in vivo* ^{31}P NMR, HPLC-UV and ^{1}H NMR metabolomics. Aquat Toxicol 76: 329–342

Viant MR, Pincetich CA, Tjeerdema RS (2006b) Metabolic effects of dinoseb, diazinon and esfenvalerate in eyed eggs and alevins of chinook salmon (*Oncorhynchus tshawytscha*) determined by ^{1}H NMR metabolomics. Aquat Toxicol 77:359–371

Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J (2005) Mass spectrometry in metabolome analysis. Mass Spectr Rev 24:613–646

Wasser JS, Lawler RG, Jackson DC (1996) Nuclear magnetic resonance spectroscopy and its applications in comparative physiology. Physiol Zool 69:1–34

Wong DCL, Whittle D, Maltby L, Warren P (2003) Multivariate analyses of invertebrate community responses to a C(12–15)AE-3S anionic surfactant in stream mesocosms. Aquat Toxicol 62:105–117

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: May 21, 2006; Accepted: November 11, 2006
Proofs received from author(s): February 6, 2007

Vol. 332: 307–310, 2007 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published March 5

OPEN ACCESS

Molecular biorepositories and biomaterials management: enhancing the value of high-throughput molecular methodologies for the natural sciences

D. L. Distel*

Ocean Genome Legacy, 240 County Road, Ipswich, Massachusetts 01938, USA

ABSTRACT: Genomics, proteomics, metabolomics and a rapidly growing list of 'omic' methodologies have the capacity to transform our view of biology and ecology by tapping the enormous information content of biomolecules. Although procedurally diverse, these approaches share common elements. All use high-throughput technologies to explore a given class of biomolecules in its totality in a given organism, species or community. All generate large quantities of data from individual and often minute samples. All hold promise to provide new insights into the ontogeny, phylogeny, physiology and ecology of organisms and their communities. However, the full potential of these new technologies is unlikely to be realized unless renewed attention is paid to description, preservation, authentication and distribution of biological source materials: issues that have long been of fundamental concern to traditional biologists but that have received less attention in recent years. This renewed emphasis on biomaterial management will require new methods and new types of biological collections (i.e. molecular biorepositories) specifically designed to address and anticipate the needs of these rapidly evolving technologies. Such biorepositories can help advance the nascent 'omics revolution' by allowing researchers to better access and preserve source materials, disseminate research products and data, share ideas, control financial and environmental costs, integrate with traditional methods and knowledge bases, and extract more meaningful data from biological specimens.

KEY WORDS: Genomic conservation · Biobanking · Biological banking · Genome banking · Genome resource repositiory · Biorepository · Biodiversity · Bioinformatics · Archiving

Resale or republication not permitted without written consent of the publisher

INTRODUCTION

As molecular methods become more important and more widely applied in the natural sciences, the focus of many biologists and ecologists has shifted from tangible physical specimens (tissues and organisms) to molecules and molecular data. Molecular data provide insights unimaginable just a few years ago, and molecular technologies allow biologists to generate and disseminate data with ever-increasing swiftness and ease. Ongoing development of high-throughput technologies for genomics, proteomics, etc. promises to further accelerate the rate at which we are able to obtain information from biological systems. However, with so much excitement surrounding the generation and analysis of molecular data, less attention is being paid to describing, verifying and archiving the biological materials that serve as the sources of these data.

If unabated, this trend could have serious consequences for the natural sciences. Traditional practices of specimen management were developed for critical practical purposes—to provide high quality, minimally altered biomaterials for research, ensure open access to these materials, preserve these materials for reanalysis and critical reevaluation, and to maintain a strong link between these materials and their environmental context. Historically, these practices have called on researchers to deposit appropriately preserved source materials in public collections (e.g. museums, botanical gardens, zoological parks, etc.) along with detailed information on the method of their collection and the sites from which they were obtained. This is particularly the

*Email: distel@oglf.org

© Inter-Research 2007 · www.int-res.com

case for those materials used for description of new taxa or previously undescribed properties of known taxa. Such biorepositories, in turn, provide a variety of functions of value to researchers.

Identification, authentication and validation

The validity of ecological and biological studies depends critically on correct identification of taxa and assignment of specimens to them. Taxonomy is a dynamic science and taxonomic designations are periodically modified to accommodate new information. For this reason, it has long been considered prudent to deposit preserved specimens (vouchers) in public collections so that taxonomic assignments can be revisited and reevaluated. Materials deposited in public collections also serve as references and subject matter for future taxonomic investigation.

Access

Natural environments of research interest are often remote and may be both difficult and expensive to access. This is especially true of marine environments. Biological materials sampled from such environments often remain in the 'private' collections of individual researchers, even when collected at public expense. The practice of depositing valuable research materials in public collections can make them available to a much broader scientific audience. This can maximize sample accessibility, minimize collecting costs and sample loss and reduce the potential environmental impacts of repeated research collection on endangered species and sensitive environments.

Policy and regulation

Biorepositories function to develop and enforce policies that ensure fair, appropriate, equitable and safe use of biomaterials. This not only provides legal, fiscal and health protections for depositors and end users, but also provides legitimacy and safeguards that ease the processes of obtaining funding and permission for collecting biomaterials in foreign states, on private or public lands and from endangered or regulated species and habitats.

Communication, cooperation and transparency

Public biorepositories increase the efficacy of research by fostering communication among researchers and by connecting authenticated specimens with field observations, experimental results and ecological data. Deposited materials also allow reinterpretation, challenge, dispute or corroboration of published observations and conclusions, thereby improving the rigor and transparency of the scientific process.

Preservation and standards

Biological specimens are perishable and the information that they contain is ephemeral. Centralized public biorepositories provide expertise, standardized methods and quality controls, specialized storage facilities and a level of storage security and continuity that cannot easily be duplicated by independent researchers.

Although modern high-throughput technologies promise fundamental changes in the ways in which research is done in marine ecology, and many other disciplines of natural science, they in no way lessen the need for appropriate specimen management. In fact the opposite is true. Such methods add considerable value to biological source materials and considerable cost to their analysis compared with more traditional methods. This creates a responsibility to ensure the proper management and use of biomaterials.

VALUE IN BIOMOLECULES

High-throughput methodologies add value to biological materials in a variety of ways. Chief among these is the capacity to extract more information from less material. Microgram or even nanogram quantities of biomaterials can often yield information from millions of base pairs of DNA or RNA, or tens of thousands of genes, proteins, and metabolites. Moreover, such tiny specimens are often called on to serve as proxies for anything from individual cells and organisms to entire communities. The resultant molecular data are easily and widely disseminated via electronic media, and can be used without modification across a broad range of research disciplines.

Molecular methods also make it possible to generate many new types of functional and informative derivatives of biomolecules, e.g. DNA or cDNA clone libraries, gene or whole genome amplification products, synthetic DNAs, microarrays and a variety of functional genes and gene products. These derivatives can be broadly distributed and used by others as primary research materials.

Additionally, molecular methods lend new value to biological samples by making it possible to preserve not only morphological information, but also to preserve, replicate and propagate biochemical information and biological functions. Thus, appropriately pre-

served biomaterials can now serve as sources of potentially valuable biochemical agents, pathways and biomarkers and as historical records of genetic and metabolic diversity, evolution, function and potential.

Finally, molecular methods add conservation value to preserved biomolecules by allowing these materials to be used to establish historical baselines for assessment of species, population and ecosystem change (Ryder et al. 2000, Ryder 2005). This may be of considerable importance to ecologists given the high rates of physical and biotic change and species and population extinction suspected in many parts of the ocean. In this respect, preservation of biomaterials may be regarded as a form of *ex situ* conservation that can contribute information valuable for the protection of threatened species and ecosystems (Ryder 2005). Beyond this, preserved biomolecules may also contribute to the resurrection of extinct genes and genomes. Indeed, complete functional genomes of extinct viruses have already been reconstructed from archival materials (Lamb & Jackson 2005), and partial genome sequences of extinct cave bears, woolly mammoths and Neanderthal man have been reconstructed from fossil tissue and bone fragments (Noonan et al. 2005, Dalton 2006, Poinar et al. 2006). Clearly, these new technologies can add extraordinary value to preserved biomaterials derived from thriving, threatened and extinct species and ecosystems.

COSTS AND BENEFITS

While genomic, proteomic and other 'omic' methods add value, they also contribute to costs of research. On the one hand, compared with traditional methods, high-throughput molecular analyses (e.g. massively parallel sequencing technologies, robotics, array-based hybridization and mass spectrometry methods) require large capital investments and typically generate high incidental and amortized costs per experiment. On the other hand, as these technologies have evolved, the unit costs of data have fallen precipitously. For example, genome sequencing costs have declined >100-fold since the start of the human genome project in 1990, and there is reasonable expectation that this trend will continue (Lander & Austin 2002). These cost decreases, however, are economies of scale that may not extend to small-scale projects and small institutions.

Given this trend, the per-datum sample analysis cost for high-throughput methods may become comparatively small relative to other research costs. However, no such cost decreases can be expected in sample collection and management. If anything, the costs of collecting biological materials, particularly in the marine realm, can be expected to increase with increasing energy costs, decreasing abundance of key taxa, increasing regulation of collection in national and international waters, and increasing need to explore more remote and inaccessible environments. Costs associated with identification and description of specimens can also be expected to rise as traditional taxonomic expertise becomes more rare.

This combination of high initial capital costs, high per-experiment costs, high costs of sample collection and limited access to marine environments will likely change the way research is conducted and funded in marine ecology, favoring large research institutions over small. By broadening access to biomaterials, data and ideas, and by providing low cost services such as DNA sequencing or library construction, molecular biorepositories can extend economy of scale to small-scale users.

In summary, high-throughput methods now make it possible to derive, disseminate and utilize more information and value from individual biological specimens than ever before, albeit often at greater total cost. Thus, these new technologies greatly increase value of individual biological specimens and so proportionately increase the potentially harmful consequences of their misidentification, incorrect documentation, mishandling or loss. The obvious conclusion is that new technologies call for increasing emphasis on appropriate specimen management rather than the decreasing trend that has been evident in recent years.

PROBLEMS AND SOLUTIONS

Surprisingly, there are few incentives for modern biologists to practice good specimen management, even as support has grown for analogous improvements in data management practices. For example, many funding agencies enforce strict policies requiring submission of sequence data, trace files, assemblies and quality evaluations to existing public data repositories as a prerequisite for funding of large-scale genomics projects. Similarly, many journals will not publish results of genomics investigations without sequence accession numbers assigned by public databases. However, few organizations enforce similar policies with regard to specimens, source materials and derivatives from genomics projects. As a result, numerous genome-sequencing projects have been completed, at considerable public expense, with no genomic DNAs or voucher specimens on deposit in public collections and little information made available describing collection, identification, locale and ecological or physiological context of source materials. A relatively small investment in proper biomaterial management could not only produce more meaningful data from such costly projects,

but could also provide the opportunity for retrospective analyses if suspicion of taxonomic misidentification or mishandling of source materials arises.

The current shift in priorities to favor data management over sample management likely has multiple causes. It may be due in part to changes in the ways that biologists are trained. New technologies have relaxed barriers that formerly separated scientific disciplines, allowing many researchers to enter new research fields without benefit of exposure to the distinct traditional practices of those fields. This problem is likely exacerbated by shortages of mentors with classical training in such areas as systematics, taxonomy, anatomy and histology. However, a more important contributing factor may be the lack of appropriate biorepositories and methods suitable to meet the developing needs of modern environmental and natural sciences. Few existing biorepositories, museums and natural history collections are equipped to provide inexpensive long-term storage of non-medical biomaterials in ways that preserve informative biomolecules. Methods for preservation, storage, propagation and distribution of biomaterials and informative biomolecules have not been adequately developed, tested and standardized. Broadly accepted policies that regulate ownership and liability for natural biomaterials, their derivatives and the intellectual property that stems from them are lacking. Standards have not been established to ensure responsible collection and trade of research biomaterials from threatened or endangered species and habitats. In short, infrastructure, practices, policies and methods in biomaterials management have failed to anticipate or to keep pace with the rapid changes occurring in biological and ecological research.

Nonetheless, the prospects for developing new standards and infrastructure for molecular biomaterial management are encouraging. Already, a large number of biological specimen repositories have emerged to fill the specialized needs of medical, veterinary, agricultural and toxicological research. Established repositories store and manage blood, fluids, tumors, cell cultures, reproductive products, embryos, cord blood, nucleic acids, proteins, metabolites and bioactive compounds. These biorepositories increase the speed and efficacy of research by providing broad access to limited biological materials, by fostering communication among researchers and by connecting authenticated biospecimens with clinical, experimental and epidemiological observations. Efforts are also afoot at museums, zoological parks, botanical gardens and private non-profit institutions to establish biomolecular resource repositories dedicated to the natural sciences. Examples include the Ambrose-Monell Cryo-collection (American Museum of Natural History, http://research.amnh.org/amcc/), the Center for Reproduction of Endangered Species (San Diego Zoo, http://cres.sandiegozoo.org/index.html), the RBG Kew Plant DNA Bank (Royal Botanic Gardens, Kew, http://rbgkew.org.uk/data/dnaBank/homepage.html) and the Ocean Genome Resource (Ocean Genome Legacy, http://oglf.org).

CONCLUSIONS

Clearly, good biomaterial management practices and appropriate biorepositories can contribute significantly to the successful exploitation of high-throughput technologies in the natural sciences. However, participation by individual researchers and research communities will be needed to develop the infrastructure, methods and scientific consensus that will make this a reality. Marine ecologists, like researchers in many fields of natural science, are only just beginning to add genomic, proteomic, metabolomic and other high-throughput methodologies to their arsenal of investigative techniques. Many are well acquainted with and appreciate the importance of sample management practices traditional to this field, and so are well positioned to recognize, advocate and contribute to the development of modernized biorepositories and biomaterial management methods. This can be done by following, promoting and teaching good specimen management practices, by utilizing existing collections and by urging funding agencies to support biorepositories and biorepository-related research through funding and policy decisions.

Acknowledgements. I thank Debra Pittman, Nate Eckborg and Yvette Luyten for helpful comments on the text. This work was supported by the Ocean Genome Legacy, a non-profit private research institution and genome resource biorepository dedicated to exploring and preserving the biological diversity of marine environments.

LITERATURE CITED

Dalton R (2006) Neanderthal DNA yields to genome foray. Nature 441:260–261

Lamb RA, Jackson D (2005) Extinct 1918 virus comes alive. Nature Med 11:1154–1156

Lander E, Austin R (2002) Sequencing and resequencing workshop summary: NHGRI large-scale sequencing workshop. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD

Noonan JP, Hofreiter M, Smith D, Priest JR and 6 others (2005) Genomic sequencing of pleistocene cave bears. Science 309:597–599

Poinar HN, Schwarz C, Qi J, Shapiro B and 9 others (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–394

Ryder OA (2005) Conservation genomics: applying whole genome studies to species conservation efforts. Cytogenet Genom Res 108:6–15

Ryder OA, McLaren A, Brenner S, Zhang YP, Benirschke K (2000) DNA banks for endangered animal species. Science 288:275–277

Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: August 14, 2006; Accepted: September 28, 2006
Proofs received from author(s): February 12, 2007