Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Microbiology

The information science of microbial ecology

Aria S Hahn[1,2], Kishori M Konwar[2,3], Stilianos Louca[4,5], Niels W Hanson[6] and Steven J Hallam[1,2,6,7]

A revolution is unfolding in microbial ecology where petabytes of 'multi-omics' data are produced using next generation sequencing and mass spectrometry platforms. This cornucopia of biological information has enormous potential to reveal the hidden metabolic powers of microbial communities in natural and engineered ecosystems. However, to realize this potential, the development of new technologies and interpretative frameworks grounded in ecological design principles are needed to overcome computational and analytical bottlenecks. Here we explore the relationship between microbial ecology and information science in the era of cloud-based computation. We consider microorganisms as individual information processing units implementing a distributed metabolic algorithm and describe developments in ecoinformatics and ubiquitous computing with the potential to eliminate bottlenecks and empower knowledge creation and translation.

Addresses
[1] Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC, Canada
[2] Koonkie, Inc., Menlo Park, CA, USA
[3] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[4] Institute of Applied Mathematics, University of British Columbia, Vancouver, BC, Canada
[5] Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada
[6] Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada V6T 1Z4
[7] ECOSCOPE Training Program, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

Corresponding author: Hallam, Steven J (shallam@mail.ubc.ca)

Current Opinion in Microbiology 2016, **31**:209–216

This review comes from a themed issue on **Environmental microbiology**

Edited by **Monica Vasquez** and **Steven Hallam**

http://dx.doi.org/10.1016/j.mib.2016.04.014

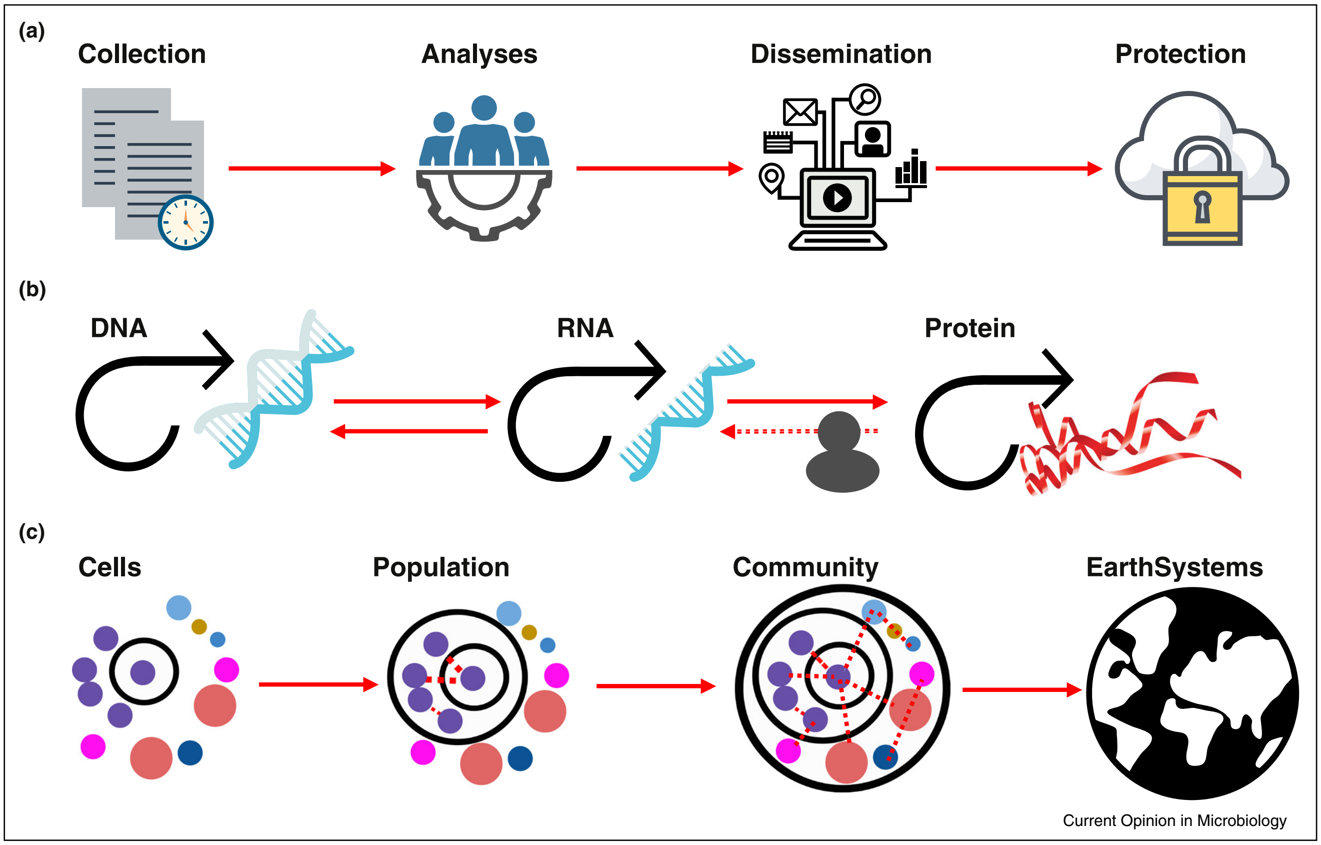
"We have to do better at producing tools to support the whole research cycle from data capture and data curation to data analysis and data visualization."
—Jim Gray, The Fourth Paradigm: Data-intensive Scientific Discovery, 2009 [1]

Introduction

Modern information theory was born in 1948 with the publication of Claude Shannon's 'The Mathematical Theory of Communication' [2•]. This work was written from the point of view of communication, and was concerned primarily with the process of correctly delivering a message from a receiver to a sender, and ignored the interpretation of the message. Fundamentally, information theory studies the limits of communication irrespective of the technologies or processes involved in the actual mechanism of message transmission, while information science leverages these theories to collect, organize, analyze, transfer and protect information (Figure 1a) [3]. Shannon described messages as consisting of a sequence of letters from a defined *alphabet*, such as the binary code used by computers, alphabets used by humans, or the genetic code of nucleic acids used by nature. A decade after Shannon's paper was published, Francis Crick proposed *the central dogma of molecular biology*, which describes the scheme of genetic information flow in biological systems [4,5] (Figure 1a,b). This work had some degree of similarity to Shannon's theory of communication. Indeed, Crick described his ideas in terms of defined *alphabets* and *information*. Still the basis of prevailing paradigm, the central dogma considers the flow of biological information, with DNA nucleotide sequences as an *alphabet*, transcribed to RNA and then translated to the amino-acid *'alphabet'*.

Here, we imagine microbes as fundamental information processing units in biology, as they are the smallest life form capable of autonomously transcribing and translating the nucleotide *alphabet* (Figure 1). Consider these information processing units, as abundant 'warehouses of entropy' implementing a distributed metabolic algorithm in a carbon-based computing cluster that works together to perform complex metabolic tasks (Figure 1c). Today, the biological information stored in microbial communities can be accessed using 'multi-omics' methods. Indeed, at present there exists an unprecedented and expanding quantity of microbial community data in the *alphabets* of DNA, RNA and protein sequences, all publicly available to the researcher. If interpretable, this cornucopia of biological information has enormous potential to reveal the metabolic networks driving matter and energy transformations in natural and engineered ecosystems, with translational benefits across a wide range of sectors including human health, biorefining and earth systems engineering. We examine microbial ecology through the lens of information science with emphasis on information transfer. We highlight current challenges in the

Figure 1



Visualization of information transmission and channels of information transmission. **(a)** Information science aims to collect, analyze, disseminate and protect information. In microbial ecology samples are collected, analyzed, published and archived. **(b)** *The central dogma of molecular biology* describes the flow of information in biological systems [4–6]. **(c)** Individual cells give rise to populations that interact to form communities of information exchange. These community interaction networks in turn help drive Earth's biogeochemical cycles.

collection, analyses, dissemination and protection of environmental sequence information and propose that cloud computing will assist and accelerate knowledge generation and scientific understanding of microorganisms on a truly global scale.

Box 1

Microbial information networks

Microbes as information processing units

Building on the conceptual model described above, we consider several similarities between microbial and computer networks. Both carbon-based and silicon-based networks contain variable metabolic or processing power, denizens that cheat [12–14] or fail, are predatory [15,16] or encode malicious processes, and both suffer from viral infection and reprogramming [17]. At present, a computer typically runs several dozens of programs, called *processes*, simultaneously, and when networked together, computers exchange messages which influence the computations carried out. Similarly, beyond the confines of laboratory environments microbes do not live in isolation, instead interacting with one another at the population and community levels to ultimately drive distributed matter and energy transformation processes (Figure 1c) [18–20]. For example, the breakdown of polysaccharides in human intestines is completed by several bacterial by multiple bacterial groups wherein the metabolic byproducts of one group serve as the primary carbon source for another group unable to degrade the original molecule [21]. Similar patterns have emerged in a biorefining context where expression of lignin transformation genes and gene cassettes encoded in different host genomes can synergize in combination to produce different monoaromatic breakdown profiles [22]. Beyond catabolic processes, microbe–host interactions require signaling processes to direct biofilm formation or differentiate host tissue structures. For example, plants actively interact with microorganisms colonizing root structures and the microbiome in turn produces signaling molecules that help shape community metabolism [23•]. Given this increasing awareness that microbial communities can work as distributed systems

Box 1 A Brief History of Information Science and Molecular Ecology

In his original work, Shannon introduced two fundamental concepts in information theory, the *source coding theorem* and the *channel coding theorem* [2•]. The source coding theorem deals with *entropy* as a fundamental unit of information and describes and quantifies the minimum rate by which a code can carry information without distorting it due to errors. The *channel coding theorem* deals with the *capacity of a channel*, which is the maximum rate at which information can be transmitted through a noisy channel that introduces errors. In 1958, the *central dogma of molecular biology* was proposed and used similar concepts to describe information transmission in biological systems. In order to build on Crick's work and access the information stored in life's complexity pyramid (DNA, RNA, protein and metabolites) [7], it became necessary to study the biological *alphabet* with increasing granularity. By 1977 Sanger sequencing allowed sequencing of DNA fragments [8]. Over the next decade, Carl Woese and Norman Pace used small subunit ribosomal RNA (SSU rRNA) genes to reveal the 'uncultured majority' of microorganisms and launched microbial ecology into the molecular era [9,10]. The advent of next-generation sequencing platforms at the end of the 20th century has resulted in an explosion of environmental sequencing projects and the generation of rich data sets in need of unification from both quantitative and comparative perspectives [11].

giving rise to ecosystem functions and services, there is growing interest in determining the role of interactions and information exchange in the structure and function of microbial communities [24–26]. While we recognize that within microbial networks the physico-chemical environment both influences and is influenced by community metabolism, the parallels between distributing work among microorganisms or processing units provides a powerful metaphor to guide the peer efforts of microbial ecologists and computer scientists in developing unifying theories and integrative software tools rooted in information theory.

Formal specification of microbial networks

Computer scientists use abstract mathematical models, such as Input Output Automata [27], to model distributed systems and computational networks to understand and analyze systems of interacting computers. Despite the complexity of biological systems, ecologists have also used network models successfully to describe different modes of species interactions and trophic structures [18,28••]. Over the past 5 years there has been marked increase in the use of network models to analyze microbial communities [28••,29–32]. Microbial network models have been used to identify keystone species that could serve as indicators for ecosystem functions [30,31], to predict protein–protein interactions that alter fundamental cellular mechanisms [33] and to determine ecological organizing principles driving spatial distribution of microbial populations within a community [32]. Despite these recent attempts to exploit the statistical properties of microbial networks, it remains difficult to validate co-occurrence patterns found in these models [28••]. Further, to date there has been limited consensus on the techniques used to build microbial networks (e.g. Spearman's correlation [29,31], Spearman's correlation and Kullback–Leibler dissimilarity measure [28••], Pearson's correlation [32,30], and ordination based co-correspondence analysis [16]) making it difficult to accurately compare network properties and build on previous work. Given these discontinuities a more formal effort to develop network standards to compare data sets within and between environments, validate hypotheses, and ultimately predict and engineer system states based on biological information transfer is needed.

Understanding information transfer in microbial community interaction networks

To chart information transfer processes in microbial community interaction networks, it becomes necessary to reconstruct compositional, regulatory and distributed metabolic processes connecting community members using multi-omic sequence information. For example, gene-centric and pathway-centric metagenomics, metatranscriptomics and metaproteomics can be used to both identify taxonomic composition and reconstruct metabolic networks on local and ecosystem scales. This information can in turn form the basis for biogeochemical models [34–38]. Similarly, initial predictions of community DNA, mRNA or protein content [39,40] can guide downstream molecular and process oriented experiments and hypothesis testing needed to validate model simulations. Ultimately, close integration of multi-omic data sets with biogeochemical parameter information and thermodynamic principles will enable time variable forecasts of microbial community metabolism and adaptive response to forcing events such as climate change [28••]. This will in turn facilitate the design of microbial communities with beneficial metabolic properties [41]. For example, just as networked computers running a distributed algorithm can complete tasks through collaboration, recent work has highlighted that the energetic burden of producing costly metabolites can be decreased through cooperation and cross-feeding [42,43••]. However, when microbial dynamics were modeled using game theory and model parameters renormalized by diffusion, a limit to cooperative benefit was identified [44•]. Interestingly, there is evidence that some microbes have evolved biological mechanisms by which to regulate the diffusion of 'public-goods' to better target interacting partners [45]. Similarly, a network of computers using distributed algorithms employs 'rumor spreading' or 'gossiping' algorithms to disseminate useful information and ensure information is effectively delivered to the target computing units. Within microbial communities these complex, but not fully understood interactions (both mutalistic and parasitic) may serve to achieve larger goals, such as enhancing a community objective function for growth and resilience [42] or delivering ecosystem services [18–20]. Comparably, large scale distributed computing systems employ complex, decentralized algorithms that effectively complete tasks while tolerating unpredictable and adversarial behavior, such as stochastic computer or network failures. On the basis of these observations, we suggest that ubiquitous computing frameworks comprised of networked processing units driven by an active user community are necessary to reconstruct and simulate microbial community interaction networks.

A future in the clouds

Information science aims to collect, analyze, disseminate and protect information. In the context of microbial ecology this requires large-scale initiatives to collect samples and biological information from natural and engineered ecosystems, analyze the collected information in a principled and reproducible manner, disseminate resulting data products and knowledge, and protect this information by making it easily available to the broader scientific community (Figure 1a). Here we discuss current efforts in academic and commercial settings to build ubiquitous computing frameworks, for example, cloud-based computing to study microbial community interaction networks.

Collecting microbial information

The increased throughput and decreased cost of next generation sequencing technologies has led to the production of vast and varied environmental data sets and large scale initiatives to make this information accessible [46–48]. Indeed, there has been a recent development for both publicly available information storage resources (e.g. the TARA ocean's project [46], the Earth Microbiome project [47], the Human Microbiome Project [48]), the National Center for Biotechnology Information [49], and the Joint Genome Institute's Genome Portal [50]). With the current average global transfer rate of 7.6 Mbps [51] even downloading such datasets could soon become a bottleneck (2.7 days). Finally, as data volumes expand, the capacity of individual hard-drives may soon be exceeded, necessitating the use of cloud computing and distributed file systems.

Enter the 'microbial' matrix

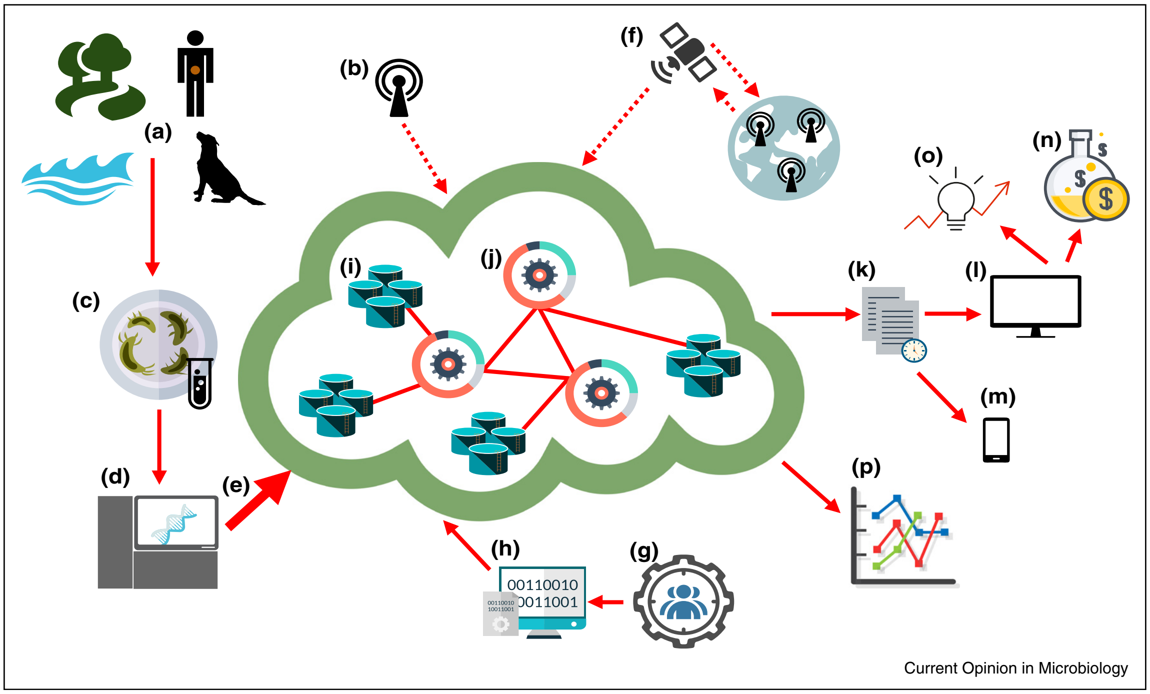
Integrating multiple sources of information including sequence reads, mass spectra and environmental parameters and processing this information using a cloud-based system can alleviate many operational and interpretive challenges faced by microbial ecologists (Figure 2). By drawing on ideas and advances from computer and information science, coders can rise to the challenges presented by the 'multi-omics' era, and design tools that scale with increasing data size, integrate software tools and frameworks more seamlessly, and empower meaningful biological insights through interactive visualization and analysis (Figure 2). For example, master-worker models can be used to process metagenomic data sets on high-performance computers in a fault tolerant manner [52], and efficient threading models can improve existing algorithms to greatly increase the speed at which data is processed [53]. Indeed, modern and centralized high performance computing (HPC) resources hold the key to overcoming persistent computational and analytic bottlenecks in the cloud.

Cloud computing gives any researcher with an Internet connection access to HPC resources obviating capital hardware investment and IT administration costs [54] (Figure 2). Consider that some of the most scalable tools, designed for parallel processing of next generation sequencing data, require upwards of 70–100 GB of computer memory (RAM) per machine to run efficiently [55•,56•]. Purchasing and maintaining these resources can place a costly hardware and administrative burden on individual researchers. Further, cloud computing reduces energy waste as the so called computational heavy lifting is done externally, requiring only inexpensive laptop and mobile devices to initiate data processing and collect results. Indeed, in many institutional settings, desktop computers utilize up to 75% of total energy consumed [57]. Finally, cloud computing can directly address reproducibility problems. Among the 'ten commandments of reproducible science', cloud-computing allows researchers to document how every result was produced, avoid manual data manipulation, archive programs, store raw data and provide public access to developed code and data products [58]. The increased throughput and decreased cost of next generation sequencing technologies enables routine generation of vast and varied environmental data sets. This trend compounded by individual processing limitations and overhead costs suggests that within 3–5 years the need for cloud computing resources will overtake conventional modes of analysis on desktop machines or local servers. Adoption of cloud-based computing is as inevitable as it is logical, as processed data can be easily pulled from the cloud and interpreted and shared on local machines, leaving the storage of raw data sets in the cloud infrastructure.

Current cloud-based services for data analysis and information storage

The popularity of cloud-based computing frameworks in business analytics, web applications and finance is rapidly driving down the cost of cloud services. However, the usage of these resources in analyzing large volume data sets in genomics is still in early stages of development [59•]. Recently, Illumina announced a centralized cloud-based service for the storage and analyses of sequence information entitled BaseSpace [60] creating an analysis ecosystem attractive to different user levels. Founded on the 'internet-of-things-' model (which describes a network of electronic devices, vehicles, and even buildings equipped with modern technologies to exchange information), Illumina's sequencers now connect directly to

Figure 2



Current Opinion in Microbiology

The matrix is a conceptual model of the interconnected network of multi-omic sequence information, processing, storage and researchers needed to chart the microcosmos. **(a)** Samples sourced from diverse natural and engineered ecosystems including our own bodies. **(c)** Samples are processed in laboratory settings **(d)** Biological information is converted into digital information via high-throughput sequencing machines, such as NGS or tandem mass spectrometers which produce petabytes of data. **(e)** Many sequencing centers can push the enormous volume of data to cloud-based storage via high-bandwidth networks. The cloud infrastructure consists of hundreds of thousands of **(i)** processing and **(j)** storage units, which collectively provide scalable data storage and processing capabilities to millions of users. **(b)** Environmental monitoring devices detect ecosystem perturbations such as harmful algal blooms or pathogenic strains in almost real time by gathering target information and transmitting to storage systems. **(f)** A network of environmental monitoring systems around the globe can collect and transmit data to storage using a multiplicity of communication links including satellites and cables. **(h)** Code for bioinformatic tools can be stored in the same infrastructure where the data resides. **(g)** Microbial ecologists, computer scientists and engineers from around the world can collaborate, refine and share their data and code. **(p)** Environmental and health professionals can gather, monitor and study the data from the monitoring sites located in far away or inaccessible places. The processed data is sent to end users via internet connections on the World Wide Web. **(l)** Desktops and **(m)** mobile devices can be used by end users to explore data interactively, while triggering on demand processing in the cloud and gathering **(k)** interpretable data, such as matrices, interactive graphics and summary statistics, to more local settings driving **(o)** idea generation and **(n)** knowledge translation.

BaseSpace, automating the uploading of sequence data output and eliminating the need for local downloads. Once on BaseSpace, sequences can be analyzed through Apps that scale Amazon Web Services (AWS) nodes to analyze data on demand. Results are then provided to the user via web-based reports. This technology makes use of other modern software advances such as Docker [61] a lightweight open source container platform that allows applications to be disseminated along with their dependencies permitting any application to be run anywhere. Similarly, Google has launched the Google Genomics could based platform, which uses distributed systems like Bigtable and Spanner to achieve a scalable and robust system with which to analyze and store petabytes of sequence data uploaded by users [62].

The National Institutes of Health (NIH) also recently launched Nephele, a cloud-computing pilot project designed to improve the efficiency and collaboration in microbiome data analysis. Nephele can be accessed online and includes several data storage repositories and pipelines for analysis of tag sequence data, for example, small subunit ribosomal gene sequences that leverage AWS. In addition to BaseSpace, Google Genomics, and Nephele, the National Foundation for Science founded a long-term cyberinfrastructure initiative entitled EarthCube whose goal is to democratize the collection, access, analyses, sharing and visualization of all forms of data and resources for geosciences in order to empower whole-earth analysis and simulation. Recognizing a need for community HPC resources, in 2013, the initiative began funding projects aimed at researching community needs and how to best design computing architecture, expanding existing conventions for metadata to promote consistent documentation, and developing Cloud-Hosted Real-time Data Services for the Geosciences (CHORDS) to manage, navigate, store, and distribute data and information via the internet [63]. Furthermore, EarthCube also initiated the EarthCube Oceanography and Geobiology Environmental Omics (ECOGEO) project which takes a multi-disciplinary approach to applying omics' technologies and bioinformatic techniques to ecological questions related to the intersection of biological, geological, and chemical processes. ECOGEO effectively integrates state-of-the-art computing platforms with microbial information and environmental data from geoscientists and oceanographers to create a community-based framework and cloud-based cyberinfrastructure for modeling and understanding Earth's systems. Indeed, such cloud-based services look to be promising models for large-scale high-throughput data processing that is reproducible between users and institutions.

Conclusion

Advances in high-throughput sequencing, computer science and distributed processing are enabling humans to perceive, reconstruct and interact with the microcosmos, a microbial matrix that defines metabolic interaction networks driving matter and energy transformations in the world around us (Figure 2). Centralized data and computing resource access points — ironically composed of thousand of distributed hardware components linked via communication channels described in Shannon's original theory of information — provide a scalable, resilient and ubiquitous system in which to access and monitor this matrix, enabling deeper insight into the ecological design principles shaping Earth systems and transforming our capacity to construct microbial solutions to vexing human problems in the years to come.

Acknowledgements

This work was carried out under the auspices of Genome Canada, Genome British Columbia, Genome Alberta, the Natural Science and Engineering Research Council (NSERC) of Canada, the Canadian Foundation for Innovation (CFI) and the Canadian Institute for Advanced Research (CIFAR) through grants awarded to SJH. ASH was supported by the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program (CGS D) administered by NSERC. KMK was supported by the Tula Foundation funded Centre for Microbial Diversity and Evolution (CMDE) at UBC. NWH and SL were supported by a four year doctoral fellowship (4YF) administered through the UBC Faculty of Graduate and Postdoctoral Studies.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hey T, Tansley S, Tolle K: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research; 2009.

2. Shannon CE: **Bell Syst Tech J**1948, **27**:379-423
• 623–656.
This seminal paper is considered as the foundation of information theory.

3. Stock WG, Stock M: *Handbook of Information Science*. Berlin/ Boston, MA: De Gruyter Saur; 2013.

4. Crick FH: **The biological replication of macromolecules**. *Symp Soc Rxp Biol*. 1958.

5. Crick FH: **Central dogma of molecular biology**. *Nature* 1970, **227**:561-563.

6. Cook ND: **The case for reverse translation**. *J Theor Biol* 1977, **64**:113-135.

7. Oltvai ZN, Barabasi AL: **Life's complexity pyramid**. *Science* 2002, **298**:763-764.

8. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**:5463-5467.

9. Woese Carl R, Fox George E: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74**:5088-5090.

10. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR: **Rapid determination of 16s ribosomal RNA sequences for phylogenetic analyses**. *Proc Natl Acad Sci U S A* 1985, **82**:6955-6959.

11. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER: **The next-generation sequencing revolution and its impact on genomics**. *Cell* 2013, **155**:27-38.

12. Morris JJ: **Black queen evolution: the role of leakiness in structuring microbial communities**. *Trends Genet* 2015, **31**:475-482.

13. Morris JJ, Papoulis SE, Lenski RE: **Coexistence of evolving bacteria stabilized by a shared black queen function**. *Evolution* 2014, **68**:2960-2971.

14. Estrela S, Morris JJ, Kerr B: **Private benefits and metabolic conflicts shape the emergence of microbial interdependencies**. *Environ Microbiol* 2015.

15. Chow CE, Winget DM III, White RA, Hallam SJ, Suttle CA: **Combining genomic sequencing methods to explore viral diversity and reveal potential virus–host interactions**. *Front Microbiol* 2015, **6**.

16. Chow CET, Kim DY, Sachdeva R, Caron DA, Fuhrman JA: **Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists**. *ISME J* 2014, **8**:816-829.

17. Hurwitz BL, Hallam SJ, Sullivan MB: **Metabolic reprogramming by viruses in the sunlit and dark ocean**. *Genome Biol* 2013, **14**:R123 http://dx.doi.org/10.1186/gb-2013-14-11-r123.

18. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum J *et al.*: **Plankton networks driving carbon export in the oligotrophic ocean**. *Nature* 2016, **02** (advance online publication).

19. Takeuchi Nobuto, Cordero Otto X, Koonin Eugene V, Kaneko Kunihiko: **Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection**. *BMC Biol* 2015, **13**:1-11.

20. Lee J, Wu J, Deng YY, Wang J, Wang C, Wang JH, Chang CQ, Dong YH, Williams P, Zhang LH: **A cell–cell communication signal integrates quorum sensing and stress response**. *Nat Chem Biol* 2013, **9**:339.

21. Rakoff-Nahoum S, Coyne MJ, Comstock LE: **An ecological network of polysaccharide utilization among human intestinal symbionts**. *Curr Biol* 2014, **24**:40-49.

22. Strachan CR, Singh R, VanInsberghe D, Ievdokymenkoa K, Budwilld K, Mohn WW, Eltis LD, Hallam SJ: **Metagenomic scaffolds enable combinatorial lignin transformation**. *Proc Natl Acad Sci U S A* 2014, **111**:10143-10148.

23. Lebeis SL, Paredes SH, Lundberg DS, Breakfield N, Gehring J,
• McDonald M, Malfatti S, del Rio TG, Jones CD, Tringe SG, Dangl JL: **Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa**. *Science* 2015, **349**:860-864.
This work demonstrates the phytohormone salicylic acid produce by some plants as a defense mechanism directly affects the microbial community that colonizes the roots using isogenic *Arabidopsis thaliana* a synthetic microbial community.

24. Zarraonaindia I, Smith DP, Gilbert JA: **Beyond the genome: community-level analysis of the microbial world**. *Biol Philos* 2013, **28**:261-282.

25. Levy R, Borenstein E: **Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules**. *Proc Natl Acad Sci U S A* 2013, **110**:12804-12809.

26. Smillie CS, Smith MB, Friedman J, Cordero OX, Lawrence AD, Alm EJ: **Ecology drives a global network of gene exchange connecting the human microbiome**. *Nature* 2011, **480**:241-244.

27. Lynch NA: *Distributed Algorithms*. Morgan Kaufmann Publishers; 1996.

28. •• Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F *et al.*: **Determinants of community structure in the global plankton interactome**. *Science* 2015, **348**.

Using both genomic and environmental data, a robust co-occurrence analysis is used to model a plankton interaction network that captures predatory and symbiotic relationships. This work not only expands current knowledge about the oceans food webs but also represents one of the first attempts to validate predicted interactions, using microscopy to validate to corroborate relationships.

29. Ma B, Wang H, Dsouza M, Lou J, He Y, Dai Z, Brookes PC, Xu J, Gilbert JA: **Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in Eastern China**. *ISME J* 2016.

30. Lupatini M, Suleiman AKA, Jacques RJS, Antoniolli ZI, de Siqueria Ferreira A, Kuramae EE, Roesch LFW: **Network topology reveals high connectance levels and few key microbial genera within soils**. *Front Environ Sci* 2014, **2**:10.

31. Williams RJ, Howe A, Hofmockel KS: **Demonstrating microbial co-occurrence pattern analyses within and between ecosystems**. *Front Microbiol* 2014, **5**.

32. Zhang Z, Geng J, Tang X, Fan H, Xu J, Wen X, Ma ZS, Shi P: **Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota**. *ISME J* 2014, **8**:881-893.

33. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A *et al.*: **The binary protein–protein interaction landscape of *Escherichia coli***. *Nat Biotechnol* 2014, **32**:285-290.

34. Ogilvie LA, Bowler LD, Caplin J, Dedi C, Diston D, Cheek E, Taylor H, Ebdon JE, Jones BV: **Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences**. *Nat Commun* 2013, **4**.

35. Hanson NW, Konwar KM, Hawley AK, Altman T, Karp PD, Hallam SJ: **Metabolic pathways for the whole community**. *BMC Genomics* 2014, **15**.

36. Hawley AK, Brewer HM, Norbeck AD, Pasa-Tolic L, Hallam SJ: **Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes**. *Proc Natl Acad Sci U S A* 2014, **111**:11395-11400.

37. Aylwarda FO, Eppleya JM, Smith JM, Chavez FP, Scholin CA, Delong EF: **Microbial community transcriptional networks are conserved in three domains at ocean basin scales**. *Proc Natl Acad Sci U S A* 2015, **112**:5443-5448.

38. Ottesen EA, Young CR, Gifford SM, Eppley JM, Marin R, Schuster AC, Scholin CA, Delong EF: **Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages**. *Science* 2014, **345**:207-212.

39. Reed Daniel C, Algar Christopher K, Huber Julie A, Dick Gregory J: **Gene-centric approach to integrating environmental genomics and biogeochemical models**. *Proc Natl Acad Sci U S A* 2014, **111**:1879-1884.

40. Reed Daniel C, Breier John A, Jiang Houshuo, Anantharaman Karthik, Klausmeier Christopher A, Toner Brandy M, Hancock Cathrine, Speer Kevin, Thurnherr Andreas M, Dick Gregory J: **Predicting the response of the deep-ocean microbiome to geochemical perturbations by hydrothermal vents**. *ISME J* 2015, **9**:1857-1869.

41. Shoaie Saeed, Karlsson Fredrik, Mardinoglu Adil, Nookaew Intawat, Bordel Sergio, Nielsen Jens: **Understanding the interactions between bacteria in the human gut through metabolic modeling**. *Sci Rep* 2013, **3**:08.

42. Dimitriu T, Lotton C, Bnard-Capelle J, Misevic D, Brown SP, Lindner AB, Taddei F: **Genetic information transfer promotes cooperation in bacteria**. *Proc Natl Acad Sci U S A* 2014, **111**:11103-11108.

43. •• Mee MT, Collins JJ, Church GM, Wang HH: **Syntrophic exchange in synthetic microbial communities**. *Proc Natl Acad Sci U S A* 2014, **111**:E2149-E2156.

Using synthetic communities of *Escherichia coli*, the authors demonstrate that cross feeding of biosynthetically costly metabolites yields synergistic growth. Next the authors suggest syntrophic interaction is a widely distributed evolutionary strategy busing through genomic comparison of over 6000 sequenced genomes.

44. • Menon R, Korolev KS: **Public good diffusion limits microbial mutualism**. *Phys Rev Lett* 2015, **114**.

This paper shows that microbial dynamics can be modeled by using game theory enhanced with parameters renormalized by diffusion. Using this model the authors show that greater sharing of metabolites reduces survival via a non-equilibrium phase transition.

45. Kmmerli R, Schiessl KT, Waldvogel T, McNeill K, Ackermann M: **Habitat structure and the evolution of diffusible siderophores in bacteria**. *Ecol Lett* 2014, **17**:1536-1544.

46. Tara Oceans Data. http://www.ebi.ac.uk/about/news/press-releases/tara-oceans-data.

47. Gilbert JA, Jansson JK, Knight R: **The earth microbiome project: successes and aspirations**. *BMC Biol* 2014, **12**.

48. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**:804-810.

49. National Institute of Health (NIH). http://www.ncbi.nlm.nih.gov/.

50. Joint Genome Institute (JGI). http://genome.jgi.doe.gov/.

51. Ookla. http://www.ookla.com/.

52. Hanson NW, Konwar KM, Wu SJ, Hallam SJ: **Metapathways v2.0: a master-worker model for environmental pathway/genome database construction on grids and clouds**. *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2014)*. 2014.

53. Kim D, Hahn AS, Wu SJ, Hanson NW, Konwar KM, Hallam SJ: **Fraggenescan+: high-throughput short-read gene prediction**. *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. 2015:1-7.

54. Fisch KM, Meissner T, Gioia L, Ducom JC, Carland TM, Loguercio S, Su AI: **Omics pipe: a community-based framework for reproducible multi-omics data analysis**. *Bioinformatics* 2015, **31**:1724-1728.

55. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nat Methods* 2015, **12**:59-60.

DIAMOND is currently the fastest available software for aligning short DNA sequencing reads to a protein reference database such as NCBI-NR. DIAMOND is about four orders of magnitude times faster than BLASTX, while reporting most of the hits that BLASTX finds, with an e-value of at most 1e−5. Of course, in the sensitive mode DIAMOND is about 3 orders of magnitude faster than BLASTX.

56. • Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biol* 2014, **15**.

Kraken is an ultrafast and highly accurate software for predicting taxonomy on short metagenomic DNA sequences. While Kraken is capable of leveraging many CPUs it is only suitable for machines with larger RAM, for example, 100 GB or more is suitable.

57. Microsoft Environment. http://www.microsoft.com/environment/IT_Energy/IT_Energy.aspx.

58. Sandve GK, Nekrutenko A, Taylor J, Hovig E: **Ten simple rules for reproducible computational research**. *PLoS Comput Biol* 2013, **9**.

59. • Kim D, Konwar M, Hanson NW, Hallam SJ: **Koonkie: an automated software tool for processing environmental sequence information using Hadoop**. *BigData/SocialInformatics/PASSAT/BioMedCom Conference, Harvard University*; *December 14–16: Conference Full Papers: 2014*.

Koonkie is a software pipeline suitable for processing metagenomic sequences to predict metabolic pathways and function using resources in the Amazon EC2 cloud and represents one the first applications of Hadoop in bioinformatics.

60. BaseSpace. https://basespace.illumina.com.

61. Docker. https://www.docker.com/.

62. Google Genomics. https://cloud.google.com/genomics/what-is-google-genomics.

63. EarthCube. http://www.earthcube.com.