PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Popa O, Oldenburg E, Ebenhöh O. 2020 From sequence to information. *Phil. Trans. R. Soc. B* **375**: 20190448.
http://dx.doi.org/10.1098/rstb.2019.0448

Accepted: 17 March 2020

One contribution of 17 to a theme issue 'Integrative research perspectives on marine conservation'.

Subject Areas:
computational biology, theoretical biology, systems biology, bioinformatics

Keywords:
data, sequence, information, entropy, genome, time-series, modelling

Author for correspondence:
Oliver Ebenhöh
e-mail: oliver.ebenhoeh@hhu.de

THE ROYAL SOCIETY PUBLISHING

From sequence to information

Ovidiu Popa[1], Ellen Oldenburg[1] and Oliver Ebenhöh[1,2]

[1]Institute of Quantitative and Theoretical Biology, and [2]Cluster of Excellence on Plant Sciences, CEPLAS, Heinrich-Heine University Düsseldorf, Germany

OP, 0000-0003-4470-0378; EO, 0000-0002-0993-9247; OE, 0000-0002-7229-7398

Today massive amounts of sequenced metagenomic and metatranscriptomic data from different ecological niches and environmental locations are available. Scientific progress depends critically on methods that allow extracting useful information from the various types of sequence data. Here, we will first discuss types of information contained in the various flavours of biological sequence data, and how this information can be interpreted to increase our scientific knowledge and understanding. We argue that a mechanistic understanding of biological systems analysed from different perspectives is required to consistently interpret experimental observations, and that this understanding is greatly facilitated by the generation and analysis of dynamic mathematical models. We conclude that, in order to construct mathematical models and to test mechanistic hypotheses, time-series data are of critical importance. We review diverse techniques to analyse time-series data and discuss various approaches by which time-series of biological sequence data have been successfully used to derive and test mechanistic hypotheses. Analysing the bottlenecks of current strategies in the extraction of knowledge and understanding from data, we conclude that combined experimental and theoretical efforts should be implemented as early as possible during the planning phase of individual experiments and scientific research projects.

This article is part of the theme issue 'Integrative research perspectives on marine conservation'.

1. Introduction

When discussing the process of generating useful information from sequences, it is helpful to agree on some basic definitions. First, we need to clarify what exactly we consider a sequence and what we understand as information. When speaking about sequences, most biologists understand a sequence found in biological macromolecules, such as the sequence of nucleotides within a DNA or RNA molecule or the sequence of amino acids within a protein. Strictly speaking, sequences are far more general and describe any set of objects (real: such as chemical compounds, or abstract: such as numbers) arranged in some sequential order. In this work, we will mostly refer to biological sequences given by the order of chemicals arranged in a sequential order within a macromolecule, but would like to stress that measurements obtained at various time points also represent a sequence, from which plenty of useful information can be extracted. Such sequences were in particular important before the advent of high-throughput technologies that allow macromolecular sequences to be read efficiently. As we will discuss, sequences of sequences, i.e. time-series of biological sequence data, are a valuable method to infer information from sequences.

While sequences are rather straightforward to define in a very general sense, it is far more challenging to capture the notion of information in a simple definition. In information theory, information—or rather the generation of information—is quantified by the information entropy (or Shannon entropy, named after Claude Shannon who introduced the concept in 1948 [1]). The concept of information entropy is highly useful to determine, for example, bounds for

lossless compression, and helps in quantifying the capacity of transmission systems to transmit data.

The difficulty is that in this theory data are inherently considered to be identical with information, and the encoding and decoding processes during communication are concerned primarily with the problem of encoding, transmitting and decoding a sequence of bits—the fundamental unit of information. The important question whether the receiver actually understands the transmitted information is not considered in this theory at all.

It is very simple to calculate the Shannon entropy of an arbitrary text, and the resulting number will tell us how randomly (or non-randomly, and thus *surprisingly*) the letters are arranged into a sequence. However, the same information (for example as contained in a user manual of a microwave or any other technical device) can be written in many languages. The Shannon entropies of all these texts may be the same, or at least very similar. But for me as a receiver it makes a great deal of difference whether the text is written in English (which I understand) or in Finnish (which I don't). This example illustrates that the information content of data, as quantified by the Shannon entropy, does not help us to predict how much useful information we can extract. It further illustrates that, in addition to the data themselves, knowledge about the decoding system (here, knowledge of a language) is required to actually make use of the information. In the following, information is interpreted as 'knowledge obtained from investigation, study, or instruction',[1] which entails that besides the pure information content also the associated decoding mechanisms are considered.

Our text is structured as follows. First, we survey which information is contained in various biological sequences, and illustrate how the information content changes when considering different levels of biological organization. We would like to note here that there exist a much higher number of biological organization levels than we address in this article, and we have selected the most fundamental ones to briefly exemplify this process. Further, we outline how information is transmitted and decoded, and discuss what kind of useful information, or *knowledge*, can be obtained from the data. We then proceed towards time-series data, which, as mentioned above, also represent a sequence containing useful information, and illustrate how new knowledge and insight are produced by different types of analysis. The multiple layers encompassing different information content are illustrated in figure 1. We conclude by suggesting that experiment and theory need to collaborate more intensely, and that this collaboration, and in particular interdisciplinary communication, need to be implemented as early as possible, during the planning and experimental design phase.

2. Information in biological sequences

All life on Earth is based on genetic sequences stored in DNA. These sequences contain key information on how to manufacture and assemble the building blocks composing an organism, how to regulate the activity of various components in response to the environment, and, most importantly, how to copy this information and transmit it to future generations. Copying information is never perfect, so information can be changed and reassembled in different combinations.

Passing the information from ancestor to descendant, or laterally between organisms, while at the same time modifying it through random mutations, inevitably led to speciation [2–4], which resulted in the enormous biodiversity on this planet. Analysing the information stored in the genetic material is a first step of a comprehensive investigation of the processes required to extract and decode biological information.

(a) DNA

Understanding information as a signal that becomes valuable after decoding by a receiver, a DNA sequence contains more informative content than the sequence of the four different nucleotides that a DNA molecule is composed of.

The order of the nucleotides within the DNA sequence reduces the information entropy. In eukaryotes for example, the genome sequence contains several types of repeated nucleotide sequences (repeats). This phenomenon results in a reduction of the DNA information entropy, as was shown in an earlier study [5]. As a beneficial result, repetitive nucleotide sequences provide genetic redundancy and gene regulation by DNA folding specificity, and are important for the synthesis of proteins with similar functions [6, p. 556]. However, at the same time the order of nucleotides increases the complexity of information storage [7]: it is responsible for the helix structure, which itself affects the robustness [8] of the double helix or the accessibility [9,10] of the DNA sequence for the interaction of organic compounds or inorganic nanomaterials [11]. For example, measuring the periodicities of 10–11 bp allows the super-coiled state of genomic DNA to be determined [12–14]. Supercoiling illustrates how sequential information stored in DNA base pairs can be translated into structural information about the DNA molecule. DNA supercoiling strongly affects DNA metabolism, has influence on the molecular evolution of the DNA [15] and is one of the most fundamental regulators of global gene expression in bacteria [16,17]. The next level of coiled DNA ordering is the specific chromosome structure, which defines almost the whole library of inherited genetic information of an organism. A disorder of this information level can cause damage to a biological system, for example the duplication of one chromosome in humans (e.g. trisomy 21) results in several health problems [18]. Information stored in the non-randomly ordered nucleotide triplets (codons) [19–21] forms the basis for the genetic code. Only this code allows DNA sequences to be scanned, decoded and interpreted by the translational machinery, to be converted into amino acids in a process that enables relocating inherited information into proteins, another set of elementary biological buildings blocks. The genetic code is perhaps the most illustrative example for the fact that yielding useful information from data always requires a functioning data decoding system. Interestingly, this information transfer from DNA to protein is highly dynamic. For example, identical proteins can be synthesized with different molecular energies if the same amino acid sequence is encoded by different codons [22].

(b) Genes

Proteins, defined by the information encoded in the DNA sequence (the gene), fulfil certain functions within a living organism. Information gathered from specific marker genes allows conclusions about evolutionary forces that are responsible for adaptation and speciation processes. For example, the most commonly used marker gene in prokaryotes is 16S ribosomal RNA (rRNA) [23]. Because this gene is considered to have an essential function, it is ubiquitous, and it exhibits

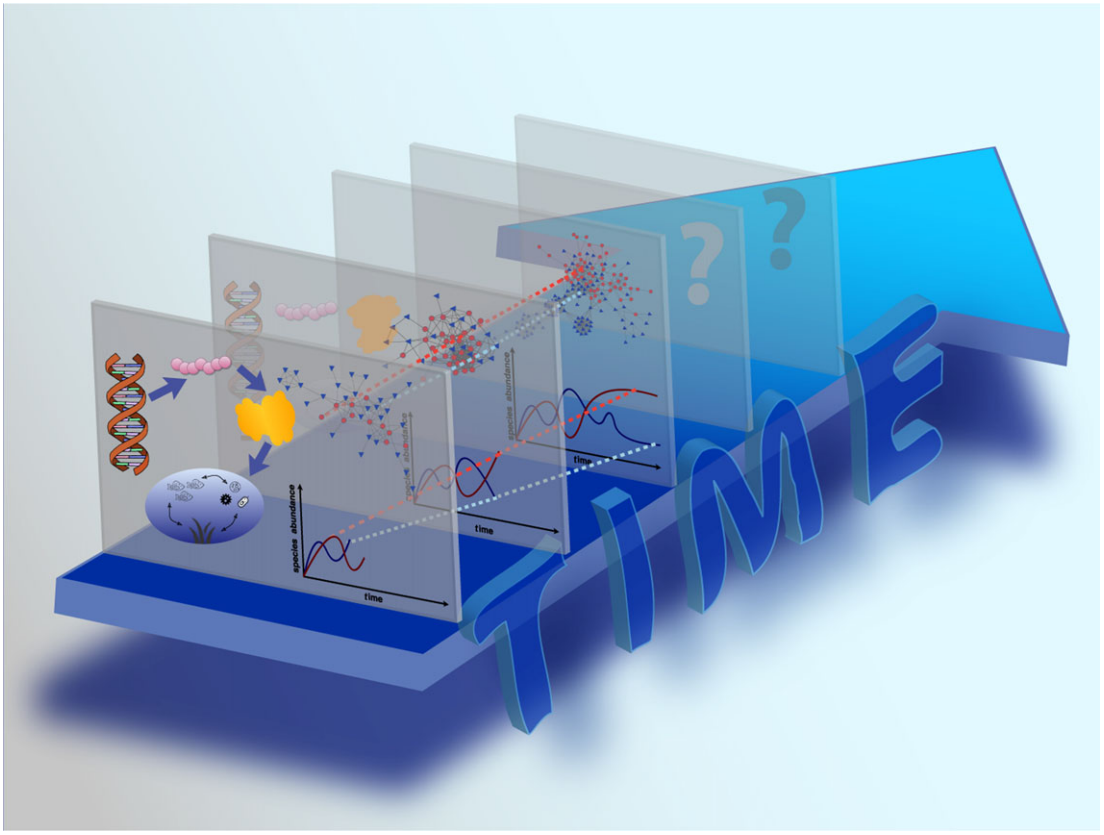

Figure 1. From sequence to information. This figure shows the different levels of information, from DNA to environment. Each layer depicts a different level of information that can be obtained from sequences. The DNA sequence encodes the genetic information that is decoded by the translational machinery into amino acid sequences. These in turn fold into functional proteins. The protein functions provide information about the capabilities of an organism such as its metabolism. Combined information of many organisms and environmental parameters characterize ecosystem dynamics. All these information layers can be used to infer different relationships, for example, in the form of networks or models. Including the temporal aspect (big blue 3D arrow), another dimension of information is gained, from which temporal correlations and interactions can be determined. A major task of time-series analysis and mechanistic modelling is to predict the future from information collected from the past. The more distant the future is that we try to predict, the more the uncertainty (question marks) increases.

a low mutation rate, comparative analyses of the DNA sequences allow reconstruction of the evolutionary history of species. Such phylogenetic reconstructions can identify clades specific to certain ecosystems, such as the SAR11 clade [24], or some archaeal species that have been identified in the euphotic zones of marine ecosystems [25]. However, interpreting results based on marker genes like 16S rRNA and thus extracting accurate information are complicated by various factors [26,27], including the experimental amplification bias, as shown by Hong *et al*. [28], or its presence in multiple copies [26]. Alternative single-copy markers like chaperonin-60 [29] or the *rpoB* gene provide more phylogenetic resolution than the 16S rRNA gene and are often used in gathering evolutionary information [23].

Proteins resulting from the translation of the DNA sequence may, in the simplest case, perform exactly one function. However, there are multiple known examples where this simple one-to-one relation is not accurate. Multifunctional proteins, the so-called 'moonlighting proteins', perform more than one biochemical or biophysical function [30,31]. Protein moonlighting means that a gene may acquire and maintain a second function without gene duplication and without loss of the primary function. As a result, such a gene is under two or more entirely different selective constraints [32]. In a nutshell, we observe that the information stored in a gene sequence is much larger than is recognized by standard comparative methods. Therefore, the optimal yield on the information stored in sequences is best obtained by the agglomeration of different research methods. Wrapping particular experimental studies in the laboratory with theoretical predictions obtained from mathematical and statistical analyses is one promising path forward to maximize the information extraction process.

(c) Genome

Zooming out from the level of single genes to the whole library of genes stored in an organism's genome allows extraction of information from the sequence in a different context. Considering the whole genome as information source, several sequence characteristics can be scanned to coax out functionality encoded in the genome structure. Focusing on the GC content variation between organisms, for example, points to genomic adaptations that might have played a significant role in the evolution of the Earth's contemporary biota [33]. In addition, genomic GC comparison allows identification of recombination events that are responsible for shaping the information flow along the genomes in an evolutionary context [34–36]. Besides the specific distribution of the nucleotides within a genome sequence, the order of genetic blocks itself entails information that is decodable and allows conclusions about mechanisms that are responsible to populate the genome with new information. Genome synteny analysis (the relative gene-order conservation

between species) can provide key insights into evolutionary chromosomal dynamics and the rearrangement rates between species [37–40]. Today, information based on complete genome sequences is mainly obtained by comparative genomic approaches. Investigation methods focusing on the pan-genome (genes present in all strains) [41] of a species elevates information mining to a new perspective. Pan-genome studies allow investigation of the plasticity of a genome at species level. For example, insertions, deletions, and recombination events, as well as single-nucleotide polymorphisms (SNPs), are only visible at pan-genome level, highlighting the consequences of evolutionary forces [41–45].

(d) Gene expression

Whereas the genomic content stored in the DNA remains rather constant throughout the lifespan of an organism, the rates with which individual genes are transcribed vary strongly over time. Transcription is regulated by multiple factors, including environmental stimuli. The result of this regulation can be observed by measuring the quantity of the messenger RNA transcripts (mRNA) under different conditions or over time. These data provide additional information that cannot be obtained from the DNA sequence alone. Transcriptomics techniques allow analysis of the entirety of all transcripts available from one organism in different tissues, under different conditions or at different time points. Information obtained by transcriptomics allows conclusions about the regulation of gene expression. There are two key contemporary techniques in the field: use of microarrays, which quantify a set of predetermined sequences, and RNA sequencing (RNA-Seq), involving high-throughput sequencing to capture all sequences [46]. For medical applications, expression data have been successful in providing a molecular basis for the diagnosis of otherwise difficult to distinguish pathologies [47–49]. In addition, co-expression profiles analysed using network and machine learning approaches [47,49,50] helped to discover functionally linked genes that are associated with specific diseases [51]. In microbiological research, co-evolutionary aspects of bacteria and their viruses (phages) are an impressive example where gene expression analysis helped in understanding the mechanistic interactions in greater detail [52,53]. Co-expression analysis is not limited to a specific genome, but can be also observed among different species, where it displays remarkably similar synchronous patterns of gene expression over time [54]. Nevertheless, the expression profile extracted and evaluated by common methods will not necessarily provide information about interactions between genes or proteins. For example, the two-component signal transduction system in bacteria is able to recognize and respond to a variety of environmental stimuli. This basic system is composed of a sensor histidine kinase that catalyses its autophosphorylation and subsequently transfers the phosphate group to a response regulator, which can then trigger different physiological changes [55–57]. This regulatory mechanism cannot be understood just by scanning the information written in the genetic sequence nor by studying its expression profile. This complex mechanism is best explored by experimental work that can be integrated in the framework of mathematical models [58–60]. Such combined interdisciplinary approaches have been successfully applied to the analysis of the MAP kinase pathway, which was performed by a combination of mathematical modelling, integrated phosphoproteomic technology and Western blotting [61]. In summary, gene expression analysis is a major contributor for our understanding of gene regulation.

(e) Functional profiling

One of the main goals of sequence analysis is the determination of functional properties. The corresponding methods are often referred to as 'functional profiling'. This process usually begins by comparative analyses of sequences of interest with annotated databases. For instance, after sequencing the protein coding gene of interest, the obtained reads are mapped on a reference database like the Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology [62–64], Clusters of Orthologous Groups (COGs) [65,66], Non-supervised Orthologous Groups (NOGs) [67], Pfam [68] and UniProt Reference (UniRef) clusters [69]. These databases are used in order to classify only protein coding sequences into a putative functional category. Efficient search methods like BLAST [70] provide a putative classification into a functional category through sequence similarity. For the assumption that similar sequences perform similar functions, this approach is very successful if a reference protein or gene/genome exists.

If the same function is encoded in highly identical protein sequences, then we would consider the information entropy of such sequences in general as very low. Sometimes sequences may perform the same function but are different in their content, e.g. in amino acid compositions. An example is the LSR2 protein, which is a transcriptional silencer found in Actinobacteria, where it binds AT-rich DNA and silences its transcription [71–73]. This example illustrates how information stored in a sequence can drastically differ depending on the level of organization that is considered: the information entropy based on the arrangement of the amino acids in the sequences is extremely high, which results from the diversity between the sequences. On the other hand, the same sequences exhibit a low information entropy when their functional properties are considered. This can be best observed when the secondary structure of the sequences is considered [71].

Functional profiling of genes and proteins is an important step in understanding the role of a sequence in the context of the whole genetic repertoire of an organism. How genes interact on the functional level is yet a higher level of information, from which new knowledge can be extracted.

(f) Pathway reconstruction

Understanding biological systems presupposes investigating how matter and energy are converted in order to maintain their functions. How exactly these processes work is very likely written in the genetic sequence. To decode it, we need more understanding than the information from sequence content alone, or how strong a gene is expressed. Rather, the interplay between various gene functions is essential. Metabolic pathway reconstruction, molecular interaction and reaction network analysis, followed by mapping processes to reference pathways, increase our understanding about a higher-level function of an organism [74–76]. Once a reaction network has been reconstructed, it can be analysed using various structural analysis techniques, such as the method of network expansion [77], or dynamic approaches based on ordinary differential equations (ODEs) [78,79]. Such approaches allow us to systematically investigate the effect of changes in

parameters that are not easily accessible experimentally, and thus to draw general conclusions about regulatory principles [80–82]. In addition, two more promising concepts for pathway analysis that assesses inherent properties in biochemical reaction networks [83,84] rely on the related concepts of elementary flux modes [85,86] and extreme pathways [87–90]. Pathway analysis undoubtedly has great potential to gain a better understanding of cellular metabolism. For example, the potential of micro-algae to uptake large quantities of phosphorus (P) and to use it as biofertilizer has been regarded as a promising way to redirect P from waste water to fields. This also makes the study of molecular mechanisms underlying P uptake and storage in micro-algae of great interest [91]. Pathway reconstruction efforts in general uncover dynamic processes that take place at cellular level and are written down in the genetic code by an evolutionary process subject to environmental adaptation pressures. Considering additional information by including environmental parameters is a necessary step towards a comprehensive understanding of the ecological processes including niche adaptation.

(g) Meta-omics: what information is there?

Fundamental research in biology heavily relies on model organisms. They have been used to uncover mechanisms that synthesize, modify, repair and degrade the genetic sequence and its encoded product, the signalling pathways that allow cells to communicate, the mechanisms that regulate gene expression and the pathways underlying diverse metabolic functions [92–96]. In order to describe many aspects of information retrieved from sequences obtained at specific time points and/or conditions, or to understand the evolutionary history of non-cultivable organisms, 'omics' data-integration techniques are essential [97,98]. Meta-omics pools the knowledge of how to read and decode the information from a sequence, as described in the previous paragraphs, together with environmental parameters that are collected with the sequences. High-throughput 'omics' techniques allow observation of metagenomes, metatranscriptomes and proteomes and thus are important to describe the behaviour of populations of uncultured microorganisms and give hints on their population genetics and biogeochemical as well as ecological interactions, which cannot easily be studied or modelled in laboratory systems [99]. High-throughput DNA sequencing enabled investigation of diverse environmental and host-associated microbial communities, thus identifying for example several new virophages [100,101] or even discovering completely new prokaryotic phyla [102]. The discovery of the Asgard superphylum, a group of uncultivated archaea including the Loki-, Thor-, Odin- and Heimdallarchaeota, and the proteins with similar features to eukaryotic coat proteins involved in vesicle biogenesis, which are present in this phylum, altered significantly our understanding of the origin of life [102,103]. These organisms were isolated from marine sediments that were sampled near Loki's Castle (a field of five hydrothermal vents that are located in the middle of the Atlantic Ocean between Greenland and Norway) [104].

Metatranscriptomics allows researchers to quantify community gene expression in an environmental sample using high-throughput sequencing technology. Today we have several pipelines (e.g. SAMSA2) to analyse the huge amount of data efficiently using high-performance computational utilities [105]. Such analyses enable quantification of gene expression and its regulation within multiple organisms in order to derive conclusions about specific molecular interactions [106]. Combining metatranscriptome and gene sequencing with time-series design allows us to collect information about the dynamics of different organisms in an environmental context [107].

Metaproteomics (community proteomics) characterizes all the proteins expressed at a given time within an ecosystem. This allows us to create hypotheses and draw conclusions about microbial functionality. Further it makes it possible to study the adaptive responses of microbes to environmental stimuli or their interactions with other organisms or host cells [108–110]. Analysis of communities in natural environments has contributed immensely to our knowledge of microbial functions, such as nutrient cycling, mutualistic endosymbiosis, organic matter degradation, metal utilization and eutrophication response [108,111–113]. Despite the additional information that is gathered through 'omics' analysis, understanding biological processes as a whole is incomplete without considering their dynamic aspects. Therefore, only by including a temporal dimension will we be able to understand and model bio-ecological processes in detail [114].

3. Ecosystem dynamics: time-series analysis

Most methods reviewed above extract and study information from genomic sequences, either alone or in a comparative context, but mostly as static structures without considering any temporal dynamics. Gene expression information describing the quantity of reads obtained either in different conditions or from different time points does contain time as a factor. Whereas comparative genomics can generate hypotheses regarding the evolutionary dynamics of genes and genomes, dynamics on shorter time-scales have not yet been discussed. It is apparent that even the best meta-omics dataset obtained for a single time point cannot yield any information regarding, for example, the mechanisms underlying the population dynamics observed in an ecosystem. Before we discuss recent and ongoing approaches to analyse time-series of sequence data and extract mechanistic information, and thus understanding, we briefly summarize essential concepts of time-series analysis in general.

The main objective of time-series modelling is to carefully collect and examine observations from the past in order to develop a suitable model that describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make predictions [115]. The prediction of time-series can therefore be described as the process of predicting the future by understanding the past.

There are many ways to analyse time-series data, depending on how much prior knowledge is available about the underlying mechanisms. Often we first distinguish between seasonal, cyclic and irregular components [116]. Analysing seasonal changes in the diversity of bacterial communities [117] has, for example, suggested that seasonal changes in environmental variables are more important than trophic interactions. Cyclic fluctuations describe recurrent medium-term changes. The metagenome data of Biller *et al.* [118] contain, for example, genomic information for a large number of bacteria, archaea, eukaryotes and viruses. The usefulness of the data is enhanced by the availability of extensive physical, chemical and biological measurements associated with each

sample. In this way, the different cyclic changes within the habitats could be investigated and possible causes identified.

When adapting a model to a dataset, particular attention should be paid to selecting the most economical model. Here, 'most economical' refers to the simplest possible model that can explain the data without overfitting [116]. One of the most popular and commonly used stochastic time-series models is the Autoregressive Integrated Moving Average (ARIMA) [119]. This model considers each time-series as a collection of linear approximation and the deviations from these fits needs to follow a statistical distribution, representing the noise. ARIMA's popularity is mainly due to the flexibility to represent several types of time-series in a simple way. There are many examples of how to use the ARIMA model [54,120–123]. An example are the effects of starfish wasting diseases in the Salian Sea, a Canadian–American border area, a marine ecosystem and global hotspot for the biodiversity of temperate asteroids with a high degree of endemism [120]. Species- and area-specific ARIMA models and their estimated parameter values showed that after the outbreak of the starfish wasting disease epidemic in 2013 the incidence of the starfish *Dermasterias imbricata* increased in three areas. The observed frequency of *D. imbricata* until 2015 exceeded the model prediction for population development. The serious limitation of the model, however, is the assumed linear form of the associated time-series, making it insufficient in many practical situations.

A commonly applied methodology for the investigation of nonlinear stochastic models is the use of artificial neural networks (ANNs). Their characteristic is the application to time-series prediction problems by their inherent ability to nonlinearly model without having to adopt the statistical distribution. The corresponding model is formed adaptively on the basis of the specified data. For this reason, ANNs are inherently data-driven and self-adaptive. The most common and popular are multi-layered perceptrons (MLPs) characterized by a single feed-forward network (FNN) with a hidden layer. This method has a wide range of applicability. For example, phage protein structures could be predicted based on the genetic sequence [124]. In a different context, the functional roles of interacting microbes could successfully be predicted from environmental parameters and intramicrobial interactions [125].

4. Mechanistic models

The strategies to analyse time-series data discussed above are essentially statistical methods that aim at extracting patterns from time-series without using prior knowledge in order to make predictions about underlying mechanisms. Mechanistic models pursue a complementary approach. Based on experimental observation and often a great deal of intuition, a researcher formulates hypotheses on certain underlying interactions that give rise to an observable macroscopic behaviour. These hypotheses are then translated into equations capturing the interactions in a quantitative way. Solving these equations generates simulation results that can be compared with experimental observations, thus verifying or falsifying the initial hypotheses. This approach has been extremely successful for relatively small systems and for very fundamental questions. Almost a century ago, Lotka [126] and Volterra [127] independently developed a simple mechanistic model of two interacting species that demonstrated how oscillations in populations of a predator and a prey species can be explained as an emergent property from simple underlying mechanistic assumptions. Not surprisingly, the Lotka–Volterra model forms the basis for a multitude of more complex models and serves as a foundation to study fundamental questions, such as the conditions for co-existence of species [128]. Generalizing the ideas and equations of Lotka and Volterra leads to the class of generalized Lotka–Volterra (gLV) models, which are commonly used to study the dynamics of ecosystems [81], including the dynamics of bacterial communities [129,130]. Whereas gLV models only contain the interacting species as variables and thus define direct interactions between species, consumer resource models developed by MacArthur [131] also consider the resources as variables. Most recently, these models have been employed to explain which environmental factors determine the species richness, i.e. the number of species that can co-exist in an ecosystem [132,133]. When the first dynamic ecosystem models were developed early during the twentieth century, no information on biological sequences was available. However, the data triggering the theories of Lotka and Volterra were time-series, i.e. sequences of estimated numbers of predator and prey species, such as the data on numbers of pelts collected by the Hudson's Bay Company [134]. Now, the question arises how time-series of biological sequence data can be employed to construct mechanistic models that generate understanding about the underlying mechanisms guiding the temporal evolution of an ecosystem.

Owing to the high throughput and the resolution, time-resolved 16S barcoding data contain information on hundreds of species. Barcoding is referred to a global bioidentification system that employs DNA sequences as unique identifiers linked mostly to a specific taxonomic unit [135]. Deriving mechanistic models from barcoding time-series was illustrated for example by Stein *et al.* [136], who developed a modified gLV model that correctly predicted the community composition of the intestinal microbiome of mice under different conditions. Based on barcoding data describing the bacterial community associated with the marine diatom *Phaeodactylum tricornutum*, Moejes *et al.* [137] demonstrated that four bacterial families dominate the phycosphere, and development of a consumer resource model illustrated the high degree of uncertainty in deriving mechanistic explanations from time-series abundance data, especially if the time resolution is low.

Genomic sequence, together with functional annotation, allows the reconstruction of genome-scale metabolic network models, which encompass the complete biochemical repertoire encoded in an organism's genome [138]. The most commonly used technique to analyse such models is flux-balance analysis (FBA) [139], which allows calculation of internal flux distributions and nutrient exchange rates for given external conditions under the assumption that the metabolism is configured in order to optimise a certain objective function, such as maximising the accumulation of biomass [140,141]. With these and related metabolic network analysis methods, such as elementary flux mode analysis [86] or the method of network expansion [77,142], it became possible for the first time to rigorously link the genotype to the phenotype, where of course the view is centred on metabolism alone [143]. Not surprisingly, the enormous power that genome-scale modelling approaches provide led to an integration of such approaches in a dynamic context. Dynamic FBA (dFBA), for example, uses the flux predictions resulting from FBA at a given time

point to dynamically update nutrient and biomass concentrations [144]. This approach was successfully employed to explain and predict the dynamics of interacting organisms and their environment [145].

The current development of modelling techniques to simulate interactions of organisms on a metabolic level proceeds with enormous momentum. Controlled mesocosm experiments [146] allow for controlled environments, in which not only the community dynamics and the temporal expression patterns can be measured, but also the micro- and macronutrients as well as cofactors in the bulk solution can be determined to derive a deeper understanding of the metabolic interdependencies within microbial communities. This clearly illustrates the key role controlled environments play in rigorously testing and improving new hypotheses and theories.

5. Conclusion

The key question for the future is how can we ensure that ongoing data collection efforts, generating vast amounts of biological sequence data, are optimally suited for the development of mechanistic models. These cannot only describe data, but also rationalize what we observe based on underlying fundamental mechanisms. It is understandable that, when a new and rather unknown system, such as the global marine microbiome, is investigated for the first time, a rather unbiased, exploratory approach is taken, as is exemplified by the *Tara* Oceans expedition [147]. The enormous mass of sequencing data is certainly useful, because it provides us with an inventory of genes that are found in marine microbes. Moreover, by combining sequence data with physical parameters and metadata, novel hypotheses can be generated, such as a functional dependence of species richness and water temperature [148]. Despite the size of the generated data resource, it still only describes a snapshot of microbial abundance, albeit with considerable detail. Thus, the information gained from the data is mostly restricted to observing what is there. It is hard to conceive that the dataset would allow answering of fundamental scientific questions, such as those regarding the underlying mechanisms guiding microbial ecosystem dynamics. It is plausible to assume that for such an endeavour a more targeted approach is required. For example, to collect barcoding, metagenome and metatranscriptome data with a high temporal and spatial resolution may be a constructive way forward towards testing specific hypotheses regarding the mechanisms by which key microbial species interact. Such experiments can be designed following some successful research guidelines from this field [149–151].

This example demonstrates that the amount of data does not necessarily correlate with the gain of basic understanding. In other examples, such as the dynamics of the phycosphere of *P. tricornutum* [137] in controlled environments, we clearly have too little data to test the numerous existing hypotheses about the mechanistic interactions between species. By discussing the information content within biological sequences and the information flow between various levels of biological organization (e.g. amino acid sequence and corresponding secondary structure), we have shown that extracting knowledge from different information layers can be more effective and fruitful for sequence data analysis. For example, in order to analyse the abundance and diversity of silencer proteins in several environments a simple comparative sequence analysis will fail owing to the high information entropy at the amino acid level. Therefore, machine learning approaches using the information stored at the different levels need to be considered as a first step of the analysis pipeline to predict putative candidates from different metagenomic datasets. In a second step, laboratory experiments need to be performed, like DNA affinity chromatography followed by ChAP-Seq analysis. This combination of computational work and laboratory experiments should highlight how important the theoretical envelope for laboratory study becomes when information from different levels is considered.

We conclude that two main aspects will become increasingly important for biological research in the near future to close the gap that currently exists between the vast amount of high-throughput data and the actual fundamental understanding generated from it. Firstly, methods need to be developed, and already existing ones need to be implemented in the daily experimental work process and refined to integrate different types of data. Today several methods already exist for particular sequence analysis [152–154]. A minority are available for time-series implementation on biological data [155–157]. This refers primarily to the integration of time-resolved sequencing data with meta-information describing external conditions like pH, temperature, dissolved oxygen, CO_2, phosphate, nitrate, salinity, pressure, chlorophyll density, etc. Moreover, novel approaches will be required to integrate results from different methods of data analysis to maximise the information gain. Secondly, after an era of mainly exploratory data acquisition, it is of paramount importance to strengthen hypothesis-driven experimental approaches [158,159]. Every research question requires its own special experimental treatment. The prevailing misconception that data acquisition comes before (and is separated from) model development often leads to a design of research projects in which interdisciplinary collaborations are restricted to the data analysis phase. In our opinion, these flaws in project design lead to inefficiency and a sub-optimal coordination between experiment and theory. As an example of how theoretical knowledge supports experimental biology, and moreover enables new insights into biological processes, we mention the studies by Marsland *et al.* and Goldford *et al.* [132,133,160]. Bridging theory and experiment, in their studies the authors monitored the assembly of hundreds of soil- and plant-derived microbiomes in well-controlled minimal synthetic media. The resulting communities were sequenced using 16S ribosomal RNA, and the outcomes were modelled mathematically. Their mathematical models could reproduce large-scale ecological patterns observed across multiple experimental settings [133].

In fact, we are convinced that the involvement of theory cannot begin too early. Bioinformaticians and modellers should be involved during experimental design, because these researchers are typically those that formulate clear working hypotheses and have a model structure in mind, even before a detailed mathematical model has been constructed. Only in close interdisciplinary discussion can the different goals and aims of experimentalists and theorists be harmonized, and experiments be planned so that the resulting data are optimally suited to build mechanistic models and test scientific hypotheses.

Data accessibility. This article has no additional data.

Authors' contributions. All authors wrote the manuscript. O.P. mainly contributed to the sections DNA, Genes, Genome, Gene expression,

Functional profiling, Pathway reconstruction and Meta-omics. E.O. mainly contributed to Ecosystem dynamics: time-series analysis. O.E. mainly contributed to Introduction, Mechanistic models and Conclusion.

Competing interests. The authors declare they have no competing interests.

Funding. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2048/1, Project ID 390686111 and the Strategischer Forschungsfond Heinrich-Heine-University Düsseldorf Project ID SFF-F 2019/1571-1 Popa.

Endnote

[1]Merriam-Webster Online Dictionary, https://www.merriam-webster.com/dictionary/information, retrieved 8 November 2019.

Reference

1. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423. (doi:10.1002/j.1538-7305.1948.tb01338.x)
2. Retchless AC, Lawrence JG. 2007 Temporal fragmentation of speciation in bacteria. *Science* **317**, 1093–1096. (doi:10.1126/science.1144876)
3. Gogarten JP, Doolittle WF, Lawrence JG. 2002 Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* **19**, 2226–2238. (doi:10.1093/oxfordjournals.molbev.a004046)
4. Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. 2009 Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* **11**, 588. (doi:10.1186/1471-2164-11-588)
5. Herzel H, Ebeling W, Schmitt AO. 1994 Entropies of biosequences: the role of repeats. *Phys. Rev. E* **50**, 5061. (doi:10.1103/PhysRevE.50.5061)
6. Solovyev VV *et al.* 1994 Black matter. In *Computer analysis of genetic macromolecules: structure, function and evolution* (eds NA Kolchanov, HA Lim), p. 513. Singapore: World Scientific. (doi:10.1142/2008)
7. Bowling JM, Bruner KL, Cmarik JL, Tibbetts C. 1991 Neighboring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Res.* **19**, 3089–3097. (doi:10.1093/nar/19.11.3089)
8. Collins M, Myers RM. 1987 Alterations in DNA helix stability due to base modifications can be evaluated using denaturing gradient gel electrophoresis. *J. Mol. Biol.* **198**, 737–744. (doi:10.1016/0022-2836(87)90214-2)
9. Kagawa TF, Stoddard D, Zhou G, Ho PS. 1989 Quantitative analysis of DNA secondary structure from solvent-accessible surfaces: the B- to Z-DNA transition as a model. *Biochemistry* **28**, 6642–6651. (doi:10.1021/bi00442a017)
10. Prinsen P, Schiessel H. 2010 Nucleosome stability and accessibility of its DNA to proteins. *Biochimie* **92**, 1722–1728. (doi:10.1016/j.biochi.2010.08.008)
11. Chen N, Li J, Song H, Chao J, Huang Q, Fan C. 2014 Physical and biochemical insights on DNA structures in artificial and living systems. *Acc. Chem. Res.* **47**, 1720–1730. (doi:10.1021/ar400324n)
12. Schieg P, Herzel H. 2004 Periodicities of 10–11 bp as indicators of the supercoiled state of genomic DNA. *J. Mol. Biol.* **343**, 891–901. (doi:10.1016/j.jmb.2004.08.068)
13. Kumar L, Futschik M, Herzel H. 2006 DNA motifs and sequence periodicities. *In Silico Biol.* **6**, 71–78.
14. Lehmann R, Machné R, Herzel H. 2014 The structural code of cyanobacterial genomes. *Nucleic Acids Res.* **42**, 8873–8883. (doi:10.1093/nar/gku641)
15. Washietl S, Machné R, Goldman N. 2008 Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.* **24**, 583–587. (doi:10.1016/j.tig.2008.09.003)
16. Sobetzko P, Travers A, Muskhelishvili G. 2012 Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proc. Natl Acad. Sci. USA* **109**, E42–E50. (doi:10.1073/pnas.1108229109)
17. Sobetzko P. 2016 Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. *Nucleic Acids Res.* **44**, 1514–1524. (doi:10.1093/nar/gkw007)
18. Mégarbané A, Ravel A, Mircher C, Sturtz F, Grattau Y, Rethoré MO, Delabar JM, Mobley WC. 2009 The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome. *Genet. Med.* **11**, 611–616. (doi:10.1097/GIM.0b013e3181b2e34c)
19. Ambrogelly A, Palioura S, Söll D. 2007 Natural expansion of the genetic code. *Nat. Chem. Biol.* **3**, 29–35. (doi:10.1038/nchembio847)
20. Woese CR. 1965 Order in the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 71–75. (doi:10.1073/pnas.54.1.71)
21. Koonin EV, Novozhilov AS. 2009 Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111. (doi:10.1002/iub.146)
22. Karafyllidis IG. 2008 Quantum mechanical model for information transfer from DNA to protein. *BioSystems* **93**, 191–198. (doi:10.1016/j.biosystems.2008.04.002)
23. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. 2007 Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73**, 278–288. (doi:10.1128/aem.01177-06)
24. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ. 2002 SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **42**, 806–810. (doi:10.1038/nature01240)
25. Delong EF. 1998 Everything in moderation: Archaea as 'non-extremophiles'. *Curr. Opin. Genet. Dev.* **8**, 649–654. (doi:10.1016/s0959-437x(98)80032-4)
26. Coenye T, Vandamme P. 2003 Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett.* **228**, 45–49. (doi:10.1016/s0378-1097(03)00717-1)
27. Roux S, Enault F, Debroas D. 2011 Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiol. Ecol.* **78**, 617–628. (doi:10.1111/j.1574-6941.2011.01190.x)
28. Hong S, Bunge J, Leslin C, Jeon S, Epstein SS. 2009 Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* **3**, 1365–1373. (doi:10.1038/ismej.2009.89)
29. Schellenberg J, Links MG, Hill JE, Dumonceaux TJ, Peters GA, Tyler S, Ball TB, Severini A, Plummer FA. 2009 Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. *Appl. Environ. Microbiol.* **75**, 2889–2898. (doi:10.1128/aem.01640-08)
30. Mani M *et al.* 2015 MoonProt: a database for proteins that are known to moonlight. *Nucleic Acids Res.* **43**, D277–D282. (doi:10.1093/nar/gku954)
31. Jeffery CJ. 1999 Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11. (doi:10.1016/S0968-0004(98)01335-8)
32. Piatigorsky J, Wistow GJ. 1989 Enzyme/crystallins: gene sharing as an evolutionary strategy. *Cell* **57**, 197–199. (doi:10.1016/0092-8674(89)90956-2)
33. Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová O. 2014 Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl Acad. Sci. USA* **111**, E4096–E4102. (doi:10.1073/pnas.1321152111)
34. Meunier J, Duret L. 2004 Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21**, 984–990. (doi:10.1093/molbev/msh070)
35. Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001 GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* **159**, 907–911.
36. Fullerton SM, Carvalho AB, Clark AG. 2001 Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142. (doi:10.1093/oxfordjournals.molbev.a003886)

37. Bhutkar A, Russo S, Smith TF, Gelbart WM. 2007 Genome-scale analysis of positionally relocated genes. *Genome Inform. Int. Conf. Genome Inform.* **17**, 1880–1887. (doi:10.1101/gr.7062307)
38. Barbazuk WB, Korf I, Kadavi C, Heyen J, Tate S, Wun E, Bedell JA, McPherson JD, Johnson SL. 2000 The syntenic relationship of the zebrafish and human genomes. *Genome Res.* **10**, 1351–1358. (doi:10.1101/gr.144700)
39. Sinha AU, Meller J. 2007 Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinf.* **8**, 82. (doi:10.1186/1471-2105-8-82)
40. Peacock CS *et al.* 2007 Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* **39**, 839–847. (doi:10.1038/ng2053)
41. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005 The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594. (doi:10.1016/j.gde.2005.09.006)
42. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C. 2011 The *Salmonella enterica* pan-genome. *Microb. Ecol.* **62**, 487. (doi:10.1007/s00248-011-9880-1)
43. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009 Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc. Natl Acad. Sci. USA* **106**, 8605–8610. (doi:10.1073/pnas.0808945106)
44. Lefébure T, Stanhope MJ. 2007 Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* **8**, R71. (doi:10.1186/gb-2007-8-5-r71)
45. Duret L, Arndt PF. 2008 The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* **4**, e1000071. (doi:10.1371/journal.pgen.1000071)
46. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. 2017 Transcriptomics technologies. *PLoS Comput. Biol.* **13**, e1005457. (doi:10.1371/journal.pcbi.1005457)
47. Carter SL, Brechbühler CM, Griffin M, Bond AT. 2004 Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**, 2242–2250. (doi:10.1093/bioinformatics/bth234)
48. Golub TR *et al.* 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537. (doi:10.1126/science.286.5439.531)
49. Shipp MA *et al.* 2002 Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8**, 68–74. (doi:10.1038/nm0102-68)
50. Langfelder P, Horvath S. 2007 Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54. (doi:10.1186/1752-0509-1-54)
51. Trossbach SV *et al.* 2019 Dysregulation of a specific immune-related network of genes biologically defines a subset of schizophrenia. *Transl. Psychiatry* **9**, 156. (doi:10.1038/s41398-019-0486-6)
52. Lindell D *et al.* 2007 Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**, 83–86. (doi:10.1038/nature06130)
53. Lemieux AA, Jeukens J, Kukavica-Ibrulj I, Fothergill JL, Boyle B, Laroche J, Tucker NP, Winstanley C, Levesque RC. 2016 Genes required for free phage production are essential for *Pseudomonas aeruginosa* chronic lung infections. *J. Infect. Dis.* **213**, 395–402. (doi:10.1093/infdis/jiv415)
54. Ottesen EA, Young CR, Eppley JM, Ryan JP, Chavez FP, Scholin CA, DeLong EF. 2013 Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc. Natl Acad. Sci. USA* **110**, E488–E497. (doi:10.1073/pnas.1222099110)
55. Hoch JA. 2000 Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.* **3**, 165–170. (doi:10.1016/S1369-5274(00)00070-9)
56. Laub MT, Goulian M. 2007 Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.* **41**, 121–145. (doi:10.1146/annurev.genet.41.042007.170548)
57. Stock AM, Robinson VL, Goudreau PN. 2000 Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215. (doi:10.1146/annurev.biochem.69.1.183)
58. Majumdar S, Pal S. 2017 Cross-species communication in bacterial world. *J. Cell Commun. Signal.* **11**, 187–190. (doi:10.1007/s12079-017-0383-9)
59. Gao R, Stock AM. 2013 Probing kinase and phosphatase activities of two-component systems *in vivo* with concentration-dependent phosphorylation profiling. *Proc. Natl Acad. Sci. USA* **110**, 672–677. (doi:10.1073/pnas.1214587110)
60. Landry BP, Palanki R, Dyulgyarov N, Hartsough LA, Tabor JJ. 2018 Phosphatase activity tunes two-component system sensor detection threshold. *Nat. Commun.* **9**, 1433. (doi:10.1038/s41467-018-03929-y)
61. Tian T, Song J. 2012 Mathematical modelling of the MAP kinase pathway using proteomic datasets. *PLoS ONE* **7**, e42230. (doi:10.1371/journal.pone.0042230)
62. Kanehisa M. 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30. (doi:10.1093/nar/28.1.27)
63. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. 2018 New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595. (doi:10.1093/nar/gky962)
64. Kanehisa M. 2019 Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951. (doi:10.1002/pro.3715)
65. Tatusov RL. 1997 A genomic perspective on protein families. *Science* **278**, 631–637. (doi:10.1126/science.278.5338.631)
66. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2014 Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269. (doi:10.1093/nar/gku1223)
67. Huerta-Cepas J *et al.* 2015 eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293. (doi:10.1093/nar/gkv1248)
68. El-Gebali S *et al.* 2018 The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432. (doi:10.1093/nar/gky995)
69. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. 2014 UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932. (doi:10.1093/bioinformatics/btu739)
70. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
71. Pfeifer E, Hünnefeld M, Popa O, Polen T, Kohlheyer D, Baumgart M, Frunzke J. 2016 Silencing of cryptic prophages in *Corynebacterium glutamicum*. *Nucleic Acids Res.* **44**, 10 117–10 131. (doi:10.1093/nar/gkw692)
72. Pfeifer E, Hünnefeld M, Popa O, Frunzke J. 2019 Impact of xenogeneic silencing on phage–host interactions. *J. Mol. Biol.* **431**, 4670–4683. (doi:10.1016/j.jmb.2019.02.011)
73. Gordon BR, Li Y, Wang L, Sintsova A, Van Bakel H, Tian S, Navarre WW, Xia B, Liu J. 2010 Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **1074**, 5154–5159. (doi:10.1073/pnas.0913551107)
74. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M. 2008 KEGG atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **36**, W423–W426. (doi:10.1093/nar/gkn282)
75. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003 Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. (doi:10.1101/gr.1239303)
76. Hu Z *et al.* 2007 VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.* **35**, W625–W632. (doi:10.1093/nar/gkm295)
77. Ebenhöh O, Handorf T, Heinrich R. 2004 Structural analysis of expanding metabolic networks. *Genome Inform.* **15**, 35–45.
78. Hubbard JH, West BH. 2013 *Differential equations: a dynamical systems approach: ordinary differential equations*, vol. 5. Berlin, Germany: Springer.
79. Eisenhammer T, Hübler A, Packard N, Kelso JS. 1991 Modeling experimental time series with ordinary differential equations. *Biol. Cybern* **65**, 107–112. (doi:10.1007/BF00202385)
80. Matuszyńska A, Saadat NP, Ebenhöh O. 2019 Balancing energy supply during photosynthesis—a theoretical perspective. *Physiol. Plant.* **166**, 392–402. (doi:10.1111/ppl.12962)
81. Succurro A, Ebenhöh O. 2018 Review and perspective on mathematical modeling of microbial ecosystems. *Biochem. Soc. Trans.* **46**, 403–412. (doi:10.1042/bst20170265)
82. Ebenhöh O, van Aalst M, Saadat NP, Nies T, Matuszyńska A. 2018 Building mathematical

models of biological systems with modelbase. *J. Open Res. Softw.* **6**, 24. (doi:10.5334/jors.236)

83. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. 2002 Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193. (doi:10.1038/nature01166)
84. Förster J, Gombert AK, Nielsen J. 2002 A functional genomics approach using metabolomics and *in silico* pathway analysis. *Biotechnol. Bioeng.* **79**, 703–712. (doi:10.1002/bit.10378)
85. Schuster S, Dandekar T, Fell DA. 1999 Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**, 53–60. (doi:10.1016/S0167-7799(98)01290-6)
86. Schuster S, Hlgetag C. 1994 On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **2**, 165–182. (doi:10.1142/S0218339094000131)
87. Klamt S, Stelling JO. 2003 Two approaches for metabolic pathway analysis? *Trends Biotechnol.* **21**, 64–69. (doi:10.1016/s0167-7799(02)00034-3)
88. Schilling CH, Schuster S, Palsson BO, Heinrich R. 1999 Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* **15**, 296–303. (doi:10.1021/bp990048k)
89. Schilling CH, Letscher D, Palsson BO. 2000 Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248. (doi:10.1006/jtbi.2000.1073)
90. Wiback SJ, Palsson BO. 2002 Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.* **83**, 808–818. (doi:10.1016/s0006-3495(02)75210-7)
91. Singh D, Nedbal L, Ebenhöh O. 2018 Modelling phosphorus uptake in microalgae. *Biochem. Soc. Trans.* **46**, 483–490. (doi:10.1042/BST20170262)
92. Fields S, Johnston M. 2005 Whither model organism research? *Science* **307**, 1885–1886. (doi:10.1126/science.1108872)
93. Botstein D, Chervitz SA, Cherry JM. 1997 Yeast as a model organism. *Science* **277**, 1259–1260. (doi:10.1126/science.277.5330.1259)
94. Harris EH. 2001 *Chlamydomonas* as a model organism. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **52**, 363–406. (doi:10.1146/annurev.arplant.52.1.363)
95. Kaletta T, Hengartner MO. 2006 Finding function in novel targets: *C. elegans* as a model organism. *Nat. Rev. Drug Discovery* **5**, 387–399. (doi:10.1038/nrd2031)
96. Briggs JP. 2002 The zebrafish: a new model organism for integrative physiology. *Am. J. Physiol. Regul., Integr. Comp. Physiol.* **282**, R3–R9. (doi:10.1152/ajpregu.00589.2001)
97. Joyce AR, Palsson B. 2006 The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* **7**, 198–210. (doi:10.1038/nrm1857)
98. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. 2013 Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666. (doi:10.1038/msb.2013.22)
99. DeLong EF. 2009 The microbial ocean from genomes to biomes. *Nature* **459**, 200–206. (doi:10.1038/nature08059)
100. Yutin N, Kapitonov VV, Koonin EV. 2015 A new family of hybrid virophages from an animal gut metagenome. *Biol. Direct* **10**, 19. (doi:10.1186/s13062-015-0054-9)
101. Zhou J, Sun D, Childers A, McDermott TR, Wang Y, Liles MR. 2015 Three novel virophage genomes discovered from Yellowstone lake metagenomes. *J. Virol.* **89**, 1278–1285. (doi:10.1128/jvi.03039-14)
102. Guy L, Ettema TJG. 2011 The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587. (doi:10.1016/j.tim.2011.09.002)
103. Zaremba-Niedzwiedzka K *et al.* 2017 Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358. (doi:10.1038/nature21031)
104. Pedersen RB *et al.* 2010 Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat. Commun.* **1**, 126. (doi:10.1038/ncomms1124)
105. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. 2018 SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinf.* **19**, 175. (doi:10.1186/s12859-018-2189-z)
106. Westreich ST, Salcedo J, Durbin-Johnson B, Smilowitz JT, Korf I, Mills DA, Barile D, Lemay DG. 2020 Fecal metatranscriptomics and glycomics suggests that bovine milk oligosaccharides are fully utilized by healthy adults. *J. Nutr. Biochem.* **79**, 108340. (doi:10.1016/j.jnutbio.2020.108340)
107. Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. 2019 Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl Acad. Sci. USA* **116**, 25 900–25 908. (doi:10.1073/pnas.1908291116)
108. Wang DZ, Kong LF, Li YY, Xie ZX. 2016 Environmental microbial community proteomics: status, challenges and perspectives. *Int. J. Mol. Sci.* **17**, 1275. (doi:10.3390/ijms17081275)
109. Wilmes P, Bond PL. 2004 The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* **6**, 911–920. (doi:10.1111/j.1462-2920.2004.00687.x)
110. Wang DZ, Xie ZX, Zhang SF. 2014 Marine metaproteomics: current status and future directions. *J. Proteomics* **97**, 27–35. (doi:10.1016/j.jprot.2013.08.024)
111. Baldrian P, López-Mondéjar R. 2014 Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods. *Appl. Microbiol. Biotechnol.* **98**, 1531–1537. (doi:10.1007/s00253-013-5457-x)
112. Glass JB, Yu H, Steele JA, Dawson KS, Sun S, Chourey K, Pan C, Hettich RL, Orphan VJ. 2014 Geochemical, metagenomic and metaproteomic insights into trace metal utilization by methane-oxidizing microbial consortia in sulphidic marine sediments. *Environ. Microbiol.* **16**, 1592–1611. (doi:10.1111/1462-2920.12314)
113. Kleiner M, Young JC, Shah M, VerBerkmoes NC, Dubilier N. 2013 Metaproteomics reveals abundant transposase expression in mutualistic endosymbionts. *MBio* **4**, e00223-13. (doi:10.1128/mbio.00223-13)
114. Bar-Joseph Z, Gitter A, Simon I. 2012 Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* **13**, 552–564. (doi:10.1038/nrg3244)
115. Raicich F, Colucci RR. 2019 A near-surface sea temperature time series from Trieste, northern Adriatic Sea (1899–2015). *Earth Syst. Sci. Data* **11**, 761–768. (doi:10.5194/essd-11-761-2019)
116. Hipel KW, McLeod AI. 1994 *Time series modelling of water resources and environmental systems*, vol. 45. Amsterdam, The Netherlands: Elsevier.
117. Gilbert JA *et al.* 2012 Defining seasonal marine microbial community dynamics. *ISME J.* **6**, 298–308. (doi:10.1038/ismej.2011.107)
118. Biller SJ *et al.* 2018 Marine microbial metagenomes sampled across space and time. *Nat. Sci. Data* **5**, 180176. (doi:10.1038/sdata.2018.176)
119. Zhang GP. 2003 Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**, 159–175. (doi:10.1016/S0925-2312(01)00702-0)
120. Montecino-Latorre D, Eisenlord ME, Turner M, Yoshioka R, Harvell CD, Pattengill-Semmens CV, Nichols JD, Gaydos JK. 2016 Devastating transboundary impacts of sea star wasting disease on subtidal asteroids. *PLoS ONE* **11**, e0163190. (doi:10.1371/journal.pone.0163190)
121. Weymann D *et al.* 2017 The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. *Mol. Genet. Genom. Med.* **5**, 251–260. (doi:10.1002/mgg3.281)
122. Pennekamp F *et al.* 2019 The intrinsic predictability of ecological time series and its potential to guide forecasting. *Ecol. Monogr.* **89**, e01359. (doi:10.1002/ecm.1359)
123. Ianiro G, Micolano R, Di Bartolo I, Scavia G, Monini M, RotaNet-Italy Study Group. 2019 Group A rotavirus surveillance before vaccine introduction in Italy, September 2014 to August 2017. *Eurosurveillance* **24**, 1800418. (doi:10.2807/1560-7917.ES.2019.24.15.1800418)
124. Seguritan V, Alves Jr N, Arnoult M, Raymond A, Lorimer D, Burgin Jr AB, Salamon P, Segall AM. 2012 Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput. Biol.* **8**, e1002657. (doi:10.1371/journal.pcbi.1002657)
125. Larsen P, Dai Y, Collart FR. 2015 Predicting bacterial community assemblages using an artificial neural network approach. In *Artificial neural networks* (ed. H Cartwright), pp. 33–43. New York, NY: Springer.
126. Lotka AJ. 1920 Analytical note on certain rhythmic relations in organic systems. *Proc. Natl Acad. Sci. USA* **6**, 410–415. (doi:10.1073/pnas.6.7.410)
127. Volterra V. 1926 Variazioni e fluttuazioni del numero d'individui in specie animali conviventi

[Variations and fluctuations of the number of individuals in cohabiting animal species]. *Mem. Acad. Lincei Roma* **2**, 31–113. (In Italian.)

128. Hofbauer J, Hutson V, Jansen W. 1987 Coexistence for systems governed by difference equations of Lotka-Volterra type. *J. Math. Biol.* **25**, 553–570. (doi:10.1007/BF00276199)
129. Mounier J, Monnet C, Vallaeys T, Arditi R, Sarthou AS, Hélias A, Irlinger F. 2008 Microbial interactions within a cheese microbial community. *Appl. Environ. Microbiol.* **74**, 172–181. (doi:10.1128/AEM.01338-07)
130. Faust K, Raes J. 2012 Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538–550. (doi:10.1038/nrmicro2832)
131. MacArthur R. 1970 Species packing and competitive equilibrium for many species. *Theor. Popul. Biol.* **1**, 1–11. (doi:10.1016/0040-5809(70)90039-0)
132. Marsland R, Cui W, Mehta P. 2019 The minimum environmental perturbation principle: a new perspective on niche theory. *arXiv*. (http://arxiv.org/abs/1901.09673)
133. Marsland R, Cui W, Goldford J, Sanchez A, Korolev K, Mehta P. 2019 Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS Comput. Biol.* **15**, e1006793. (doi:10.1371/journal.pcbi.1006793)
134. Hewitt CG. 1921 *The conservation of the wild life of Canada*. New York, NY: C. Scribner.
135. Hebert PD, Cywinska A, Ball SL, DeWaard JR. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218)
136. Stein RR, Bucci V, Toussaint NC, Buffie CG, Rätsch G, Pamer EG, Sander C, Xavier JB. 2013 Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.* **9**, e1003388. (doi:10.1371/journal.pcbi.1003388)
137. Moejes F, Succurro A, Popa O, Maguire J, Ebenhöh O. 2017 Dynamics of the bacterial community associated with *Phaeodactylum tricornutum* cultures. *Processes* **5**, 77. (doi:10.3390/pr5040077)
138. Fell DA, Poolman MG, Gevorgyan A. 2010 Building and analysing genome-scale metabolic models. *Biochem. Soc. Trans.* **38**, 1197–1201. (doi:10.1042/BST0381197)
139. Orth JD, Thiele I, Palsson BO. 2010 What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248. (doi:10.1038/nbt.1614)
140. Schuetz R, Kuepfer L, Sauer U. 2007 Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119. (doi:10.1038/msb4100162)
141. Schuster S, Pfeiffer T, Fell DA. 2008 Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.* **252**, 497–504. (doi:10.1016/j.jtbi.2007.12.008)
142. Handorf T, Ebenhöh O, Heinrich R. 2005 Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512. (doi:10.1007/s00239-005-0027-1)
143. Ebenhöh O, Handorf T. 2009 Functional classification of genome-scale metabolic networks. *EURASIP J. Bioinform. Syst. Biol.* **2009**, 570456. (doi:10.1155/2009/570456)
144. Mahadevan R, Edwards JS, Doyle FJ. 2002 Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* **83**, 1331–1340. (doi:10.1016/S0006-3495(02)73903-9)
145. Perez-Garcia O, Lear G, Singhal N. 2016 Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Front. Microbiol.* **7**, 673. (doi:10.3389/fmicb.2016.00673)
146. Fahimipour AK, Hein AM. 2014 The dynamics of assembling food webs. *Ecol. Lett.* **17**, 606–613. (doi:10.1111/ele.12264)
147. Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. 2015 *Tara* Oceans studies plankton at planetary scale. *Science* **348**, 873–873. (doi:10.1126/science.aac5605)
148. Sunagawa S *et al.* 2015 Structure and function of the global ocean microbiome. *Science* **348**, 1261359. (doi:10.1126/science.1261359)
149. Faust K, Lahti L, Gonze D, de Vos WM, Raes J. 2015 Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25**, 56–66. (doi:10.1016/j.mib.2015.04.004)
150. Nason GP, Powell B, Elliott D, Smith PA. 2017 Should we sample a time series more frequently? Decision support via multirate spectrum estimation. *J. R. Stat. Soc. A* **180**, 353–407. (doi:10.1111/rssa.12210)
151. Marx CJ. 2013 Can you sequence ecology? Metagenomics of adaptive diversification. *PLoS Biol.* **11**, e1001487. (doi:10.1371/journal.pbio.1001487)
152. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016 Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977. (doi:10.1371/journal.pcbi.1004977)
153. Wirbel J. 2019 Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689. (doi:10.1038/s41591-019-0406-6)
154. Ignacio-Espinoza JC, Ahlgren NA, Fuhrman JA. 2019 Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat. Microbiol.* **5**, 265–271. (doi:10.1038/s41564-019-0628-x)
155. Ridenhour BJ, Brooker SL, Williams JE, Van Leuven JT, Miller AW, Dearing MD, Remien CH. 2017 Modeling time-series data from microbial communities. *ISME J.* **11**, 2526–2537. (doi:10.1038/ismej.2017.107)
156. Klevecz RR, Murray DB. 2001 Genome wide oscillations in expression – wavelet analysis of time series data from yeast expression arrays uncovers the dynamic architecture of phenotype. *Mol. Biol. Rep.* **28**, 73–82. (doi:10.1023/A:1017909012215)
157. Shade A, Caporaso JG, Handelsman J, Knight R, Fierer N. 2013 A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J.* **7**, 1493. (doi:10.1038/ismej.2013.54)
158. Kerr MK, Churchill GA. 2001 Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201. (doi:10.1093/biostatistics/2.2.183)
159. Auer PL, Doerge R. 2010 Statistical design and analysis of RNA sequencing data. *Genetics* **185**, 405–416. (doi:10.1534/genetics.110.114983)
160. Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, Segrè D, Mehta P, Sanchez A. 2018 Emergent simplicity in microbial community assembly. *Science* **361**, 469–474. (doi:10.1126/science.aat1168)