 doi:10.20944/preprints201911.0252.v1

From Sequence to Information

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

[1]Institute of Quantitative and Theoretical Biology, Heinrich-Heine University Düsseldorf

[2]Cluster of Excellence on Plant Sciences, CEPLAS

Abstract: Today massive amounts of sequenced metagenomic and transcriptomic 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 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 is of critical importance. We review diverse techniques to analyse time-series data and discuss various approaches by which time-series of biological sequence data was successfully used to de-rive 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.

Keywords: data; sequence; information; entropy; genome; gene; proteins; time-series; modeling; meta-genomics; transcriptomics; proteomics; bioinformatics; DNA

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 nucleobases 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,

 doi:10.20944/preprints201911.0252.v1

but would like to stress that also measurement values obtained at various time points 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 allowing to read macromolecular sequences efficiently. As we will discuss, sequences of sequences, i. e. time series of biological sequence data, are the most interesting and informative type of data to analyse in the context of extracting information from sequences.

While sequences are rather straight-forward 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 [115]). Shannon himself considered a model of data communication, where data, generated by a sender, is transmitted through a communication channel and the receiver is faced with the problem to identify the original data generated by the sender [115]. The concept of information entropy is highly useful to determine, for example, bounds for lossless compression and helps quantifying the capacity of transmission systems to transmit data. The difficulty is that in this theory data is inherently considered to be identical with information, and the encoding and decoding processes during communication are concerned primarily with the problem to encode, transmit and decode a sequence of bits – the fundamental unit of information. The important question whether the receiver actually understands the transmitted information is not considered, or rather, assumed to be always the case. It is very simple to calculate the Shannon entropy of an arbitrary text, and the resulting number will tell us how random (or non-random, and thus *surprising*) the letters are arranged into a sequence. This gives us a glimpse of how much information is contained in the text. 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 itself, 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 as quantified by the Shannon entropy, also the associated decoding mechanisms are considered.

Our text is structured as follows. First, we survey which information is contained in various biological sequences, illustrate how the information content changes when considering different levels of biological organisation, outline how informa-

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

 doi:10.20944/preprints201911.0252.v1

tion 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 represents a sequence containing useful information, and illustrate how new knowledge and insight is produced by different types of analysis. Subsequently, we review recent approaches to extract understanding from time-series of biological sequence data. The multiple layers encompassing different information content is illustrated in Figure 1. Our observations include that often the amount of data may be enormous, but the insights that can be obtained from it, remains rather limited. We conclude by suggesting that experiment and theory need to collaborate more intensely, and that this collaboration and in particular the interdisciplinary communication, needs to be implemented as early as possible already during the planning and experimental design phase. A close interaction and exchange of ideas before the actual data is obtained stands a great chance to increase the potential knowledge extracted from the data tremendously.

2 Information in biological sequences

All life on earth is based on genetic sequences stored in the DNA. This sequence contains key information on how to manufacture and assemble the building blocks comprising 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. Sequences that store information that can efficiently be used to copy the sequences have a selective advantage over sequences in which this information is less useful. Thus, it is essentially the usefulness of information which is subject to evolutionary pressures [89, 1, 100, 139, 64]. 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 [104, 37, 12] that 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.

2.1 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 actual information content depends on how the signal is interpreted. For example, on the most basic level each nucleotide (A,C,G,T) carries information of a particular biochemical property. Thus, each nucleotide has a highly specific base-pairing partner that allows creating a stable structure capable to persistently store genetic information. The order of the nucleotides within the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 November 2019 doi:10.20944/preprints201911.0252.v1

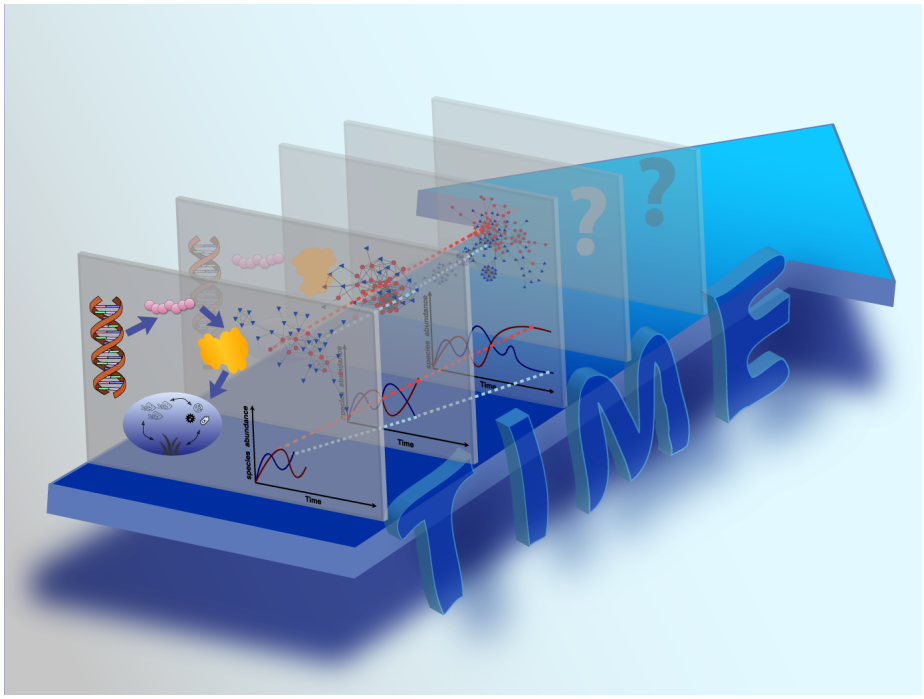

Figure 1: From sequence to information. This figure shows the different levels of information, from DNA to environment. Each layer depicts the different level of information that can be obtained from sequences. The DNA sequence encodes the genetic information that is decoded by the translational machinery 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 characterise 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 (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 mark) increases.

 doi:10.20944/preprints201911.0252.v1

DNA sequence reduces the information entropy of that sequence. However, at the same time the complexity of information storage is increased [11]: the order of nucleotides is responsible for the helix structure, which itself affects the robustness [18] of the double helix or the accessibility [57, 101] of the DNA sequence for the interaction of organic compounds or inorganic nanomaterials [16]. For example measuring the periodicities of 10-11 bp allow determining the super-coiled state of genomic DNA [107, 65, 71]. 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 and the three-dimensional structure of DNA is essential for its function [53, 4] and has influence on the molecular evolution of genomic DNA [135]. Further, DNA supercoiling is one of the most fundamental regulators of global gene expression in bacteria [122, 121], and was often shown to affect bacterial growth and transcription regulation [36, 132, 131]. Information stored in the non-randomly ordered nucleotide triplets (codons) [3, 138, 64] 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. The data stored in a coding region of the DNA is only useful in conjunction with ribosomes decoding the nucleotide sequences, translating it into a sequence of proteins. 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 [62].

2.2 Genes

Proteins, defined by the information encoded in the DNA sequence (the gene) fulfil certain functions within a living organism. Comparative analyses therefore attempt to deduce specific functionalities, and thus to extrapolate information stored in the DNA exploiting previous knowledge. For example, information gathered from specific marker genes allows for conclusions about evolutionary forces that are responsible for adaptation and speciation processes. The most commonly used marker gene is the 16S rRNA gene in prokaryotes [15]. Because this gene is considered to have an essential function, it is ubiquitous, and it exhibits a low mutation rate, comparative analyses of the DNA sequences allow reconstructing the evolutionary history of species. Such phylogenetic reconstructions gave rise to the existence of new clades specific to certain ecosystems, such as the SAR11 clade [87] or some archaeal species that have been identified in the euphotic zones of marine ecosystems [19]. However, interpreting results based on marker genes like 16S rRNA and thus extracting accurate information is complicated by various factors [17, 105], including the experimental amplification bias, as shown by Hong et al. 2009 [49], or its presence in multiple copies [17]. Alternative single copy markers like chaperonin-60 [106] or the rpoB gene

 doi:10.20944/preprints201911.0252.v1

provided more phylogenetic resolution than the 16S rRNA gene and are often used in gathering evolutionary information [15].
What exactly the information stored in the gene sequence tells us about the function of that gene in the context of a whole organism, is far from clear. 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, so-called "moonlighting proteins", perform more than one biochemical or biophysical function [78, 55]. 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, these genes are under two or more entirely different selective constraints [99]. In a nutshell, we observe that the information stored in a gene sequence is much larger as is recognized by standard comparative methods. Therefore, the optimal grazing on the information stored in sequences is best obtained by the agglomeration of different research methods. Wrapping experimental studies in the lab with theoretical predictions obtained from mathematical and statistical analyses is one promising path forward to maximize the information yield.

2.3 Genome

Zooming out from the level of single genes to the whole library of genes sorted in an organism's genome, allows extracting 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. In a comparative analysis of whole genomes, different aspects of encoded information can be obtained. Focusing on the GC content variation between organisms, for example, points to genomic adaptations associated with changing GC content that might have played a significant role in the evolution of the Earth's contemporary biota [120]. In addition, genomic GC comparison allows identifying recombination events that are responsible to shape the information flow along the genomes in an evolutionary context. [83, 33, 31]. Besides the specific distribution of the nucleotides within a genome sequence, the order of genetic blocks itself entails information that is decodeable and allows for conclusions on 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. [7, 6, 119, 93]. Today the main bunch of information we dig out of the whole sequence is due to comparative genomic approaches. Classical organism relation and evolutionary related studies are therefore performed with the objective to understand which loci in the genome was shaped by mutations, what function is encoded there and how is it linked to adaptation or environmental conditions? Investigation methods focusing on the pan genome (genes present in all strains) [82] of a species elevates the information mining space into a new perspective. Living the single genome level new questions arises that often targets raising the knowledge of how a

 doi:10.20944/preprints201911.0252.v1

species can be best described. The pan genome studies allows to investigate the plasticity of a genome on species level. For example insertion, deletions, recombination events as well as single nucleotide polymorphisms (SNPs) become first visible on pan genome level highlighting the consequences of evolutionary forces [82, 54, 103, 70, 21].

2.4 Gene expression

Information mining from sequences contains understanding its composition, the ability to form different influential structures and knowing what the product of its content harbours after decoding by the translational machinery. Whereas the genomic content stored in the DNA remains rather constant throughout the life span 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 transcripts (mRNA) under different conditions or over time. This data provides additional information that cannot be obtained from the DNA sequence alone.

For medical applications, expression data have been successful in predicting the class of unknown samples, providing a molecular basis for the diagnosis of otherwise difficult to distinguish pathologies [14, 38, 117]. In addition co-expression profile analysed using network and machine learning approaches [67, 14, 117] allowed already to discover functional linked genes that are associated to e.g. a specific diseases [133]. In microbiological research, co-evolutionary aspects of bacteria and their viruses (phages) is an impressive example where gene expression analysis helped understanding the mechanistic interactions in greater detail [73, 72].

Co-expression analysis is not limited to a specific genome, it can be also observed among related populations, as well as very different species where it displays remarkably similar synchronous patterns of gene expression over time.[92]

Nevertheless the expression profile extracted and evaluated by common methods will not necessary provide information about interaction between genes or proteins. So for example the two-component signal transduction systems 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 catalyzes its autophosphorylation and then subsequently transfers the phosphate group to a response regulator, which can then trigger different physiological changes. [47, 69, 125]. This regulatory mechanism cannot be understood just by scanning the information written in the genetic sequence and neither by exploring its expression profile. This complex mechanism is best explored by experimental work that can be integrated in the framework of mathematical models [77, 34, 66].

Gene expression analysis is a major contributor for our understanding of which putative function a particular gene encodes for.

 doi:10.20944/preprints201911.0252.v1

2.5 Functional profiling

One of the main goal of sequence analysis is the determination of functional properties. This process usually begins by comparative analysis of sequence of interest with annotated databases. Typically after sequencing the gene or genome of interests, the obtain reads are mapped on a reference database like Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology [59, 61, 60], Clusters of Orthologous Groups (COGs) [130, 32], Non-supervised Orthologous Groups (NOGs) [51], Pfam [25] and UniProt Reference (UniRef) Clusters [129]. Methods like BLAST [2] search are fast and immediately provide a putative classification into a functional category through sequence similarity. However, this approach introduces a high degree of uncertainty and assumes that related, previously sequenced organisms contain a similar set of genes with a recognizable sequence similarity. Sometimes sequences may perform same function but are different in their content, e.g. in amino acid compositions. Like the LSR2 protein, which is a transcription silencer found in Actiobacteria, in which it binds AT reach DNA and silence its transcription [98, 97, 39]. This example illustrates how information stored in the sequence can drastically differ between the different levels of observations. 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 under inspection. Functional profiling of genes or proteins of interest is an important step in understanding the role of the sequence inside 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.

2.6 Pathway reconstruction

Understanding biological systems presupposes investigating how matter and energy are converted in order to maintain its function. How exactly these processes work is very likely written in the genetic sequence. To decode it we need more understanding then the information from sequence content, or how strong a gene is expressed. Rather, the interplay between various gene functions is essential. Metabolic pathway reconstruction, molecular interaction and reaction networks analysis, followed by mapping processes to reference pathways, increases our understanding about a higher-level function of an organism [90, 116, 50]. Once a reaction network has been reconstructed, it can be analysed using various structural analysis techniques, such as the method of network expansion [23], or dynamic approaches based on ordinary differential equations (ODEs). Such approaches allow systematically investigating the effect of changes in parameters, which are not easily accessible experimentally, and thus drawing general conclusions about regulatory principles [81, 126, 24]. In addition, two more promising concepts for pathway analysis that assesses inherent properties in biochemical reaction networks [124, 30], rely on the related concepts of elementary flux modes [111, 112] and extreme pathways [63, 109, 108, 137]. Pathway

 doi:10.20944/preprints201911.0252.v1

analysis undoubtedly has great potential to gain a better understanding of the cellular metabolism. For example the potential of micro-algae to uptake large quantities of phosphorus (P) and to use it as biofertiliser has been regarded as a promising way to redirect P from waste water to the field. This also makes the study of molecular mechanisms underlying P uptake and storage in micro-algae of great interest [118]. Pathway reconstruction efforts in general uncovers dynamical processes that take place on cellular level and are written down in the genetic code by an evolutionary process subject to environmental adaptation pressure. Considering additional information by including environmental parameters is a necessary step towards a comprehensive understanding of the ecological processes including niche adaptation.

2.7 MetaOmics - What information is there?

Fundamental research in biology heavily relies on model organisms. They allowed to uncover mechanisms that synthesize, modify, repair, and degrade the genetic sequence and its encoded product, the signaling pathways that allow cells to communicate, the mechanisms that regulate gene expression and the pathways underlying diverse metabolic functions. [29, 10, 44, 58, 13]. This abundance of information yield unprecedented views of the cellular inner workings. In order to describe many aspects of the information retrieved from sequences regarding specific time points and/or conditions, or to understand the evolutionary history of non-cultivable organisms, omics data-integration techniques are essential [56]. High-throughput "omics" techniques allow observing metagenomes, -transcriptomes 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 [20]. Scanning the information from a whole bunch of sequences that is typically collected from a particular environment enables the opportunity of advancing major discoveries, significantly. High-throughput DNA sequencing enabled investigating diverse environmental and host-associated microbial communities, thus identifying for example several new virophages [140, 142] or even complete new prokaryotic phyla [41]. The discovery of the Asgard superphylum, a group of uncultivated archaea including the Lokiarchaeota, 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 [41, 141]. These organisms were identified 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) [94]. Despite the additional auxiliary information that is gathered through omics analysis which usually relies on a snapshot from the target environment, understanding biological processes as a whole needs to consider their dynamic aspect as well. herefore, only by including a temporal dimension we will be able to understand and model bio-ecological processes in detail [5].

 doi:10.20944/preprints201911.0252.v1

3 Eco-system dynamics – Time series analysis

Most methods reviewed above study and extract information from genomic sequences, either standing alone or in a comparative context, but as a static structure without considering any temporal dynamics. 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 will discuss recent and ongoing approaches to analyse time-series of sequence data and extract mechanistic information, and thus understanding, we briefly summarise essential concepts of time-series analysis in general.

The main objective of time series modeling 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 [102]. 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 analyze time series data, depending on how much prior knowledge is available about the underlying mechanisms. Often one first distinguishes between seasonal, cyclic and irregular components [46]. Analysing seasonal changes in the diversity of bacterial communities [35] 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. [8] contain for example genomic information of 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 data set, particular attention should be paid to selecting the most economical model. Here, economic refers to the simplest possible model that can explain the data without overfitting [46]. One of the most popular and commonly used stochastic time series models is the Autoregressive Integrated Moving Average (ARIMA). The basic assumption that these models are based on is that the time series under consideration is piecewise linear and that deviations from this behaviour follow a statistical distribution representing noise. The 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 this model [86, 136, 95, 52, 92]. 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 [86]. 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 *D. imbricata* increased in

 doi:10.20944/preprints201911.0252.v1

3 areas. The observed frequency of *D. imbricata* until 2015 exceeded the model prediction for population development. The serious limitation of these models, however, is the assumed linear form of the associated time series, making them insufficient in many practical situations.
A commonly applied methodology for the investigation of nonlinear stochastic models are 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 [114]. In a different context, the functional roles of interacting microbes could successfully be predicted from environmental parameters and intramicrobial interactions [68].

4 Mechanistic models

The strategies to analyse time series discussed above are essentially statistical methods that aim at extracting patterns from time series without using prior knowledge. While machine learning and artificial intelligent approaches may be useful to detect patterns that humans might miss, and display remarkable successes in making predictions [84, 40] it is highly challenging to extract useful information about underlying mechanisms from these models. Mechanistic models pursue the opposite 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, usually with computers using numerical methods, generates simulation results that can be compared to 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 [74] and Volterra [134] 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 only does this model explain a huge class of observations (the exact system where oscillation occurs is irrelevant for the mathematical formulation), but the underlying terms defining interactions between species possess a simple form that can easily be adapted and generalised to new systems for which more detailed data is available. 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 [48]. Generalising the ideas and equations of Lotka and Volterra leads to the class

 doi:10.20944/preprints201911.0252.v1

of generalised Lotka-Volterra (gLV) models, which are commonly used to study the dynamics of ecosystems [126], including the dynamics of bacterial communities [88, 27]. Whereas gLV models only contain the interacting species as variables and thus define direct interactions between species, consumer resouce models developed by MacArthur [75] also consider the resources as variables. Most recently, these models are employed to explain which environmental factors determine the species richness, i. e. the number of species that can co-exist in an ecosystem [80, 79].Apparently, when the first dynamic ecosystem models were developed early during the 20^{th} century, no information on biological sequences were 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 captured by the Hudson's Bay Company [45]. 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. Species abundance data, as approximated by 16S barcoding sequences for bacterial communities, are a straight-forward modern variant of the data used by Lotka and Volterra. However, due to the high throughput and the resolution, time resolved 16S barcoding data always contains information on hundreds of species. This illustrates that a higher data resolution does not always lead to a better understanding of the data, but may in fact divert the focus from key mechanisms to side effects not relevant for the principle dynamic behaviour. Nevertheless, successes have been achieved to derive mechanistic models from barcoding time-series, as illustrated for example by Stein et al. [123], who could develop 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 *Phaedactylum tricornutum*, Moejes et al. [85] could demonstrate 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.

Obviously, the existing data is much richer than abundance information based on barcoding sequences alone. The availability of high quality fully sequenced genomes, in combination with efficient functional annotation tools to infer the functions of the individual genes, was a major breakthrough for the theoretical sciences. 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 [28]. Analysing these models therefore allows studying the biochemical potential within a huge variety of organisms. The most commonly used technique to analyse such models is Flux-Balance Analysis (FBA) [91], which allows calculating 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 [110, 113]. With these and related metabolic network analysis meth-

 doi:10.20944/preprints201911.0252.v1

ods, such as Elementary Model Analysis [112] or the Method of Network Expansion [23, 42], it became possible for the first time to rigorously link the genotype to the phenotype, where of course the view is centered on metabolism alone [22]. Still, in no other field of biology have we advanced this far in our quest to predict phenotypic traits from genotypic sequence information alone, as in biochemistry.

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 [76]. This approach was successfully employed to explain and predict the dynamics of interacting organisms and their environment [96]. Later, a spatial component was added [43], and effects such as 'metabolic shading' of populations could be correctly predicted. Generalising the dFBA approach to allow for the dynamic description of regulatory circuits allowed to hypothesise that observed changes of substrate preference in *Escherichia coli* populations emerge from the dynamics of different sub-populations [127]. The current development of modelling techniques to simulate interaction of organisms on a metabolic level proceeds with enormous momentum. It can be expected that the inclusion of metatranscriptome data will be highly useful to assess which parts of metabolism are active at which time, and eventually lead to novel predictions how metabolism is regulated within a dynamic ecosystem environment. Moreover, controlled mesocosm experiments [26] 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 on the metabolic interdependencies within microbial communities. This clearly illustrates the key role controlled environments play to rigorously test and improve 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 can not only describe data, but also rationalise 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 [9]. 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, together with physical parameters and metadata, novel hypotheses can be generated, such as a functional dependence of species richness and water temperature [128]. Despite the size of the generated data resource, it still only describes a snapshot of microbial abundance, albeit with consid-

 doi:10.20944/preprints201911.0252.v1

erable detail. Thus, the information gained from the data is mostly restricted to observing what is there. It is questionable, though, whether the entirety of Tara data will help to gain information in the sense of obtaining knowledge and understanding. For example, it is hard to conceive that the dataset would allow answering 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 for a small number of selected locations 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. Moreover, it would be possible to dissect the local temporal dynamics from spatial dynamics introduced through drifts and currents.

This example demonstrates that the amount of data does not necessarily correlate with the gain of basic understanding. In some cases, such as for Tara Oceans, we may actually be confronted with far more data than we can comprehend with our current understanding of biology. In other examples, such as the dynamics of the phycosphere of *Phaeodactylum tricornutum* [85] 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 organisation, we have shown that extracting information from sequences is most effectively done by combining several different methods of data analysis. For example, functional annotation of gene sequences provide us with information on metabolism, whereas regularities of 10-11 bp repeats gives information on the three dimensional structure and thus provides hints on genetic regulatory mechanisms. On the other hand, it became apparent that to answer a particular scientific question and to test a specific hypothesis requires an appropriate collection of data, tailored to the specific problem at hand.

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 refined to integrate different types of data. This refers primarily to the integration of time-resolved sequencing data with meta-information describing external conditions. 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. Every research question requires its own special experimental treatment. The pervailing 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. In fact, we are convinced that the involve-

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

Contribution of Authors
OP mainly contributed to the sections DNA, Genes, Genome, Gene expression, Functional profiling, Pathway reconstruction and MetaOmics. EO mainly contributed to Eco-system dynamics – Time series analysis. OE mainly contributed to Introduction, Mechanistic models and Conclusion. All authors wrote the manuscript.

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

References

[1] Sophie S Abby, Eric Tannier, Manolo Gouy, and Vincent Daubin. Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 2012.

[2] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[3] Alexandre Ambrogelly, Sotiria Palioura, and Dieter Söll. Natural expansion of the genetic code, 2007.

[4] L Aravind, Roman L Tatusov, Yuri I Wolf, D Roland Walker, and Eugene V Koonin. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends in Genetics*, 14(11):442–444, 1998.

[5] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. Studying and modelling dynamic biological processes using time-series gene expression data, 2012.

[6] W. Bradley Barbazuk, Ian Korf, Candy Kadavi, Joshua Heyen, Stephanie Tate, Edmund Wun, Joseph A. Bedell, John D. McPherson, and Stephen L. Johnson. The syntenic relationship of the zebrafish and human genomes, 2000.

[7] Arjun Bhutkar, Susan Russo, Temple F. Smith, and William M. Gelbart. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome informatics. International Conference on Genome Informatics*, 2006.

[8] Steven J Biller, Paul M Berube, Keven Dooley, Madeline Williams, Brandon M Satinsky, Thomas Hackl, Shane L Hogle, Allison Coe, Kristin Bergauer, Heather A Bouman, Thomas J Browning, Daniele de Corte, Christel Hassler, Debbie Hulston, Jeremy E Jacquot, Elizabeth W Maas, Thomas Reinthaler, Eva Sintes, Taichi Yokokawa, and Sallie W Chisholm. Marine microbial metagenomes sampled across space and time. *Nature Scientific Data*, 5, 2018.

[9] P. Bork, C. Bowler, C. de Vargas, G. Gorsky, E. Karsenti, and P. Wincker. Tara oceans studies plankton at planetary scale. *Science*, 348(6237):873–873, may 2015.

[10] David Botstein, Steven A. Chervitz, and J. Michael Cherry. Yeast as a model organism, 1997.

[11] John M. Bowling, Kaylon L. Bruner, Joan L. Cmarik, and Clark Tibbetts. Neighboring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Research*, 1991.

[12] Marit S Bratlie, Jostein Johansen, Brad T Sherman, Da Wei Huang, Richard A Lempicki, and Finn Drabløs. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics*, 11(1):588, 2009.

[13] Josephine P. Briggs. The zebrafish: A new model organism for integrative physiology, 2002.

[14] Scott L. Carter, Christian M. Brechbühler, Michael Griffin, and Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 2004.

[15] Rebecca J. Case, Yan Boucher, Ingela Dahllöf, Carola Holmström, W. Ford Doolittle, and Staffan Kjelleberg. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, 2007.

[16] Nan Chen, Jiang Li, Haiyun Song, Jie Chao, Qing Huang, and Chunhai Fan. Physical and biochemical insights on DNA structures in artificial and living systems. *Accounts of Chemical Research*, 2014.

[17] Tom Coenye and Peter Vandamme. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 2003.

[18] Mary Collins and Richard M. Myers. Alterations in DNA helix stability due to base modifications can be evaluated using denaturing gradient gel electrophoresis. *Journal of Molecular Biology*, 1987.

[19] Edward F. Delong. Everything in moderation: Archaea as 'non-extremophiles'. *Current Opinion in Genetics and Development*, 1998.

[20] Edward F. DeLong. The microbial ocean from genomes to biomes, 2009.

[21] Laurent Duret and Peter F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, 2008.

[22] Oliver Ebenhöh and Thomas Handorf. Functional classification of genome-scale metabolic networks. *EURASIP J Bioinform Syst Biol*, 2009.

[23] Oliver Ebenhöh, Thomas Handorf, and Reinhart Heinrich. Structural analysis of expanding metabolic networks. *Genome Inform*, 15(1):35–45, 2004.

[24] Oliver Ebenhöh, Marvin van Aalst, Nima P. Saadat, Tim Nies, and Anna Matuszyńska. Building Mathematical Models of Biological Systems with modelbase. *Journal of Open Research Software*, 2018.

[25] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432, 2018.

[26] Ashkaan K. Fahimipour and Andrew M. Hein. The dynamics of assembling food webs. *Ecol Lett*, 17(5):606–613, May 2014.

[27] Karoline Faust and Jeroen Raes. Microbial interactions: From networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012.

[28] David A Fell, Mark G Poolman, and Albert Gevorgyan. Building and analysing genome-scale metabolic models. *Biochem Soc Trans*, 38(5):1197–1201, Oct 2010.

[29] Stanley Fields and Mark Johnston. Whither model organism research?, 2005.

[30] Jochen Förster, Andreas Karoly Gombert, and Jens Nielsen. A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology and Bioengineering*, 2002.

[31] Stephanie M. Fullerton, Antonio Bernardo Carvalho, and Andrew G. Clark. Local rates of recombination are positively correlated with GC content in the human genome, 2001.

[32] Michael Y. Galperin, Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1), nov 2014.

[33] N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis, 2001.

[34] Rong Gao and Ann M. Stock. Probing kinase and phosphatase activities of two-component systems in vivo with concentration-dependent phosphorylation profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.

[35] Jack A. Gilbert, Joshua A. Steele, J. Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, Alice C. McHardy, Rob Knight, Ian Joint, Paul Somerfield, Jed A. Fuhrman, and Dawn Field. Defining seasonal marine microbial community dynamics. *ISME Journal*, 6(2):298–308, 2012.

[36] Nick Gilbert and James Allan. Supercoiling in dna and chromatin, 2014.

[37] J Peter Gogarten, W Ford Doolittle, and Jeffrey G Lawrence. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19(12):2226–2238, 2002.

[38] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 1999.

[39] Blair R.G. Gordon, Yifei Li, Linru Wang, Anna Sintsova, Harm Van Bakel, Songhai Tian, William Wiley Navarre, Bin Xia, and Jun Liu. Lsr2 is a nucleoid-associated protein that targets AT-rich sequences and virulence genes in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 2010.

[40] Marco Grzegorczyk, Andrej Aderhold, and Dirk Husmeier. Overview and evaluation of recent methods for statistical inference of gene regulatory networks from time series data. In *Methods in Molecular Biology*, pages 49–94. Springer New York, dec 2018.

[41] Lionel Guy and Thijs J G Ettema. The archaeal 'TACK' superphylum and the origin of eukaryotes, 2011.

[42] T Handorf, O Ebenhöh, and R Heinrich. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *Journal of Molecular Evolution*, 61(4):498–512, 2005.

[43] William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, and et al. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7(4):1104–1115, May 2014.

[44] Elizabeth H Harris. CHLAMYDOMONAS AS A MODEL ORGANISM. *Annu. Rev. Plant Physiol. Plant Mol. Biol*, 2001.

[45] Charles Gordon Hewitt. *The conservation of the wild life of Canada*. New York: C. Scribner, 1921.

[46] Keith W Hipel and A Ian McLeod. *Time series modelling of water resources and environmental systems*, volume 45. Elsevier, 1994.

[47] James A. Hoch. Two-component and phosphorelay signal transduction, 2000.

[48] J Hofbauer, V Hutson, and W Jansen. Coexistence for systems governed by difference equations of lotka-volterra type. *Journal of mathematical biology*, 25(5):553–570, 1987.

[49] Sunhee Hong, John Bunge, Chesley Leslin, Sunok Jeon, and Slava S. Epstein. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME Journal*, 2009.

[50] Zhenjun Hu, David M. Ng, Takuji Yamada, Chunnuan Chen, Shuichi Kawashima, Joe Mellor, Bolan Linghu, Minoru Kanehisa, Joshua M. Stuart, and Charles DeLisi. VisANT 3.0: New modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Research*, 2007.

[51] Jaime Huerta-Cepas, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, Daniel R. Mende, Shinichi Sunagawa, Michael Kuhn, and et al. eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1), 2015.

[52] Giovanni Ianiro, Roberto Micolano, Ilaria Di Bartolo, Gaia Scavia, Marina Monini, RotaNet-Italy Study Group, et al. Group a rotavirus surveillance before vaccine introduction in italy, september 2014 to august 2017. *Eurosurveillance*, 24(15), 2019.

[53] Rossitza N. Irobalieva, Jonathan M. Fogg, Daniel J. Catanese, Thana Sutthibutpong, Muyuan Chen, Anna K. Barker, Steven J. Ludtke, Sarah A. Harris, Michael F. Schmid, Wah Chiu, and Lynn Zechiedrich. Structural diversity of supercoiled DNA. *Nature Communications*, 2015.

[54] Annika Jacobsen, Rene S. Hendriksen, Frank M. Aaresturp, David W. Ussery, and Carsten Friis. The Salmonella enterica Pan-genome, 2011.

[55] Constance J. Jeffery. Moonlighting proteins, 1999.

[56] Andrew R. Joyce and Bernhard Palsson. The model organism as a system: Integrating 'omics' data sets, 2006.

[57] Todd F. Kagawa, Donald Stoddard, Guangwen Zhou, and Pui S. Ho. Quantitative Analysis of DNA Secondary Structure from Solvent-Accessible Surfaces: The B- to Z-DNA Transition as a Model. *Biochemistry*, 1989.

[58] Titus Kaletta and Michael O. Hengartner. Finding function in novel targets: C. elegans as a model organism, 2006.

[59] M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, jan 2000.

[60] Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, sep 2019.

[61] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595, oct 2018.

[62] Ioannis G. Karafyllidis. Quantum mechanical model for information transfer from DNA to protein. *BioSystems*, 2008.

[63] Steffen Klamt and Jörg O. Stelling. Two approaches for metabolic pathway analysis? *Trends in Biotechnology*, 2003.

[64] Eugene V. Koonin and Artem S. Novozhilov. Origin and evolution of the genetic code: The universal enigma, 2009.

[65] Lokesh Kumar, Matthias Futschik, and Hanspeter Herzel. DNA motifs and sequence periodicities. *In Silico Biology*, 2006.

[66] Brian P. Landry, Rohan Palanki, Nikola Dyulgyarov, Lucas A. Hartsough, and Jeffrey J. Tabor. Phosphatase activity tunes two-component system sensor detection threshold. *Nature Communications*, 2018.

[67] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1:54, 2007.

[68] Peter Larsen, Yang Dai, and Frank R Collart. Predicting bacterial community assemblages using an artificial neural network approach. In *Artificial Neural Networks*, pages 33–43. Springer, 2015.

[69] Michael T. Laub and Mark Goulian. Specificity in Two-Component Signal Transduction Pathways. *Annual Review of Genetics*, 2007.

[70] Tristan Lefébure and Michael J. Stanhope. Evolution of the core and pan-genome of Streptococcus: Positive selection, recombination, and genome composition. *Genome Biology*, 2007.

[71] Robert Lehmann, Rainer Machné, and Hanspeter Herzel. The structural code of cyanobacterial genomes. *Nucleic Acids Research*, 2014.

[72] Andrée Ann Lemieux, Julie Jeukens, Irena Kukavica-Ibrulj, Joanne L. Fothergill, Brian Boyle, Jérôme Laroche, Nicholas P. Tucker, Craig Winstanley, and Roger C. Levesque. Genes required for free phage production are essential for pseudomonas aeruginosa chronic lung infections. *Journal of Infectious Diseases*, 2016.

[73] Debbie Lindell, Jacob D. Jaffe, Maureen L. Coleman, Matthias E. Futschik, Ilka M. Axmann, Trent Rector, Gregory Kettler, Matthew B. Sullivan, Robert Steen, Wolfgang R. Hess, George M. Church, and Sallie W. Chisholm. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, 2007.

[74] A J Lotka. Analytical Note on Certain Rhythmic Relations in Organic Systems. *Proc Natl Acad Sci U S A*, 6(7):410–415, jul 1920.

[75] Robert MacArthur. Species packing and competitive equilibrium for many species. *Theoretical population biology*, 1(1):1–11, 1970.

[76] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophys J*, 83(3):1331–1340, Sep 2002.

[77] Sarangam Majumdar and Sukla Pal. Cross- species communication in bacterial world, 2017.

[78] Mathew Mani, Chang Chen, Vaishak Amblee, Haipeng Liu, Tanu Mathur, Grant Zwicke, Shadi Zabad, Bansi Patel, Jagravi Thakkar, and Constance J. Jeffery. MoonProt: A database for proteins that are known to moonlight. *Nucleic Acids Research*, 2015.

[79] Robert Marsland, Wenping Cui, Joshua Goldford, Alvaro Sanchez, Kirill Korolev, and Pankaj Mehta. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLOS Computational Biology*, 15(2), feb 2019.

[80] Robert Marsland, Wenping Cui, and Pankaj Mehta. The minimum environmental perturbation principle: A new perspective on niche theory. jan 2019.

[81] Anna Matuszyńska, Nima P. Saadat, and Oliver Ebenhöh. Balancing energy supply during photosynthesis – a theoretical perspective. *Physiologia Plantarum*, 2019.

[82] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome, 2005.

[83] Julien Meunier and Laurent Duret. Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution*, 2004.

[84] Keiichi Mochida, Satoru Koda, Komaki Inoue, and Ryuei Nishii. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. *Frontiers in Plant Science*, 9, nov 2018.

[85] Fiona Moejes, Antonella Succurro, Ovidiu Popa, Julie Maguire, and Oliver Ebenhöh. Dynamics of the Bacterial Community Associated with Phaeodactylum tricornutum Cultures. *Processes*, 5(4):77, 2017.

[86] Diego Montecino-Latorre, Morgan E Eisenlord, Margaret Turner, Reyn Yoshioka, C Drew Harvell, Christy V Pattengill-Semmens, Janna D Nichols, and Joseph K Gaydos. Devastating transboundary impacts of sea star wasting disease on subtidal asteroids. *PloS one*, 11(10), 2016.

[87] Robert M. Morris, Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, William A. Siebold, Craig A. Carlson, and Stephen J. Giovannoni. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, 2002.

[88] Jérôme Mounier, Christophe Monnet, Tatiana Vallaeys, Roger Arditi, Anne Sophie Sarthou, Arnaud Hélias, and Françoise Irlinger. Microbial interactions within a cheese microbial community. *Applied and Environmental Microbiology*, 74(1):172–181, 2008.

[89] H Ochman, J G Lawrence, and E A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[90] Shujiro Okuda, Takuji Yamada, Masami Hamajima, Masumi Itoh, Toshiaki Katayama, Peer Bork, Susumu Goto, and Minoru Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research*, 2008.

[91] Jeffrey D. Orth, Ines Thiele, and Bernhard ø. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248, Mar 2010.

[92] Elizabeth A Ottesen, Curtis R Young, John M Eppley, John P Ryan, Francisco P Chavez, Christopher A Scholin, and Edward F DeLong. Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proceedings of the National Academy of Sciences*, 110(6):E488–E497, 2013.

[93] Christopher S. Peacock, Kathy Seeger, David Harris, Lee Murphy, Jeronimo C. Ruiz, Michael A. Quail, Nick Peters, Ellen Adlem, Adrian Tivey, Martin Aslett, Arnaud Kerhornou, Alasdair Ivens, Audrey Fraser, Marie Adele Rajandream, Tim Carver, Halina Norbertczak, Tracey Chillingworth, Zahra Hance, Kay Jagels, Sharon Moule, Doug Ormond, Simon Rutter, Rob Squares, Sally Whitehead, Ester Rabbinowitsch, Claire Arrowsmith, Brian White, Scott Thurston, Frédéric Bringaud, Sandra L. Baldauf, Adam Faulconbridge, Daniel Jeffares, Daniel P. Depledge, Samuel O. Oyola, James D. Hilley, Loislene O. Brito, Luiz R.O. Tosi, Barclay Barrell, Angela K. Cruz, Jeremy C. Mottram, Deborah F. Smith, and Matthew Berriman. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nature Genetics*, 2007.

[94] Rolf B. Pedersen, Hans Tore Rapp, Ingunn H. Thorseth, Marvin D. Lilley, Fernando J.A.S. Barriga, Tamara Baumberger, Kristin Flesland, Rita Fonseca, Gretchen L. Früh-Green, and Steffen L. Jorgensen. Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nature Communications*, 2010.

[95] Frank Pennekamp, Alison Iles, Joshua Garland, Georgina Brennan, Ulrich Brose, Ursula Gaedke, Ute Jacob, Pavel Kratina, Blake Matthews, Stephan Munch, et al. The intrinsic predictability of ecological time series and its potential to guide forecasting. *bioRxiv*, 2018.

[96] Octavio Perez-Garcia, Gavin Lear, and Naresh Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in Microbiology*, 7, May 2016.

[97] Eugen Pfeifer, Max Hünnefeld, Ovidiu Popa, and Julia Frunzke. Impact of Xenogeneic Silencing on Phage-Host Interactions. *Journal of molecular biology*, 2019.

[98] Eugen Pfeifer, Max Hünnefeld, Ovidiu Popa, Tino Polen, Dietrich Kohlheyer, Meike Baumgart, and Julia Frunzke. Silencing of cryptic prophages in Corynebacterium glutamicum. *Nucleic Acids Research*, 2016.

[99] Joram Piatigorsky and Graeme J. Wistow. Enzyme/crystallins: Gene sharing as an evolutionary strategy, 1989.

[100] Ovidiu Popa and Tal Dagan. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*, pages 1–9, 2011.

[101] Peter Prinsen and Helmut Schiessel. Nucleosome stability and accessibility of its DNA to proteins. *Enfermedades Infecciosas y Microbiologia Clinica*, 2010.

[102] Fabio Raicich and Renato R. Colucci. A near-surface sea temperature time series from Trieste, northern Adriatic Sea (1899-2015). *Earth System Science Data*, 11(2):761–768, 2019.

[103] Michael L. Reno, Nicole L. Held, Christopher J. Fields, Patricia V. Burke, and Rachel J. Whitaker. Biogeography of the Sulfolobus islandicus pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, 2009.

[104] A C Retchless and J G Lawrence. Temporal Fragmentation of Speciation in Bacteria. *Science*, 317(5841):1093–1096, 2007.

[105] Simon Roux, François Enault, Gisè le Bronner, and Didier Debroas. Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiology Ecology*, 2011.

[106] John Schellenberg, Matthew G. Links, Janet E. Hill, Tim J. Dumonceaux, Geoffrey A. Peters, Shaun Tyler, T. Blake Ball, Alberto Severini, and Francis A. Plummer. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. *Applied and Environmental Microbiology*, 2009.

[107] Patrick Schieg and Hanspeter Herzel. Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA. *Journal of Molecular Biology*, 2004.

[108] Christophe H. Schilling, David Letscher, and Bernhard O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 2000.

[109] Christophe H. Schilling, Stefan Schuster, Bernhard O. Palsson, and Reinhart Heinrich. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress*, 1999.

[110] Robert Schuetz, Lars Kuepfer, and Uwe Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli. *Molecular Systems Biology*, 3, jul 2007.

[111] Stefan Schuster, Thomas Dandekar, and David A. Fell. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering, 1999.

[112] Stefan Schuster and Claus Hlgetag. On elementary flux modes in biochemical reaction systems at steady state. 2(2):165–182, 1994.

[113] Stefan Schuster, Thomas Pfeiffer, and David A. Fell. Is maximization of molar yield in metabolic networks favoured by evolution? *Journal of Theoretical Biology*, 252(3):497–504, jun 2008.

[114] Victor Seguritan, Nelson Alves Jr, Michael Arnoult, Amy Raymond, Don Lorimer, Alex B Burgin Jr, Peter Salamon, and Anca M Segall. Artificial neural networks trained to detect viral and phage structural proteins. *PLoS computational biology*, 8(8), 2012.

[115] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[116] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[117] Margaret A. Shipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Ricardo C.T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 2002.

[118] Dipali Singh, Ladislav Nedbal, and Oliver Ebenhöh. Modelling phosphorus uptake in microalgae, 2018.

[119] Amit U. Sinha and Jaroslaw Meller. Cinteny: Flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 2007.

[120] Petr Šmarda, Petr Bureš, Lucie Horová, Ilia J. Leitch, Ladislav Mucina, Ettore Pacini, Lubomír Tichý, Vít Grulich, and Olga Rotreklová. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[121] Patrick Sobetzko. Transcription-coupled DNA supercoiling dictates the chromosomal arrangement of bacterial genes. *Nucleic Acids Research*, 2016.

[122] Patrick Sobetzko, Andrew Travers, and Georgi Muskhelishvili. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[123] Richard R. Stein, Vanni Bucci, Nora C. Toussaint, Charlie G. Buffie, Gunnar Rätsch, Eric G. Pamer, Chris Sander, and João B. Xavier. Ecological

modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLOS Computational Biology*, 9(12):1–11, 12 2013.

[124] Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 2002.

[125] Ann M. Stock, Victoria L. Robinson, and Paul N. Goudreau. Two-Component Signal Transduction. *Annual Review of Biochemistry*, 2000.

[126] Antonella Succurro and Oliver Ebenhöh. Review and perspective on mathematical modeling of microbial ecosystems. *Biochemical Society Transactions*, mar 2018.

[127] Antonella Succurro, Daniel Segrè, and Oliver Ebenhöh. Emergent subpopulation behavior uncovered with a community dynamic metabolic model of escherichia coli diauxic growth. *mSystems*, 4(1), jan 2019.

[128] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, and Others. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.

[129] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu and. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, nov 2014.

[130] R L Tatusov. A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637, 1997.

[131] Andrew Travers and Georgi Muskhelishvili. A common topology for bacterial and eukaryotic transcription initiation? *EMBO Reports*, 2007.

[132] Andrew Travers, Georgi Muskhelishvili, Tony Kouzarides, and Andrew J Bannister. Bacterial chromatin This review comes from a themed issue on Differentiation and gene regulation Edited. *Current Opinion in Genetics & Development*, 2005.

[133] Svenja V Trossbach, Laura Hecher, David Schafflick, René Deenen, Ovidiu Popa, Tobias Lautwein, Sarah Tschirner, Karl Köhrer, Karin Fehsel, Irina Papazova, Berend Malchow, Alkomiet Hasan, Georg Winterer, Andrea Schmitt, Gerd Meyer Zu Hörste, Peter Falkai, and Carsten Korth. Dysregulation of a specific immune-related network of genes biologically defines a subset of schizophrenia. *Translational psychiatry*, 9(1):156, 2019.

[134] V Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei Roma*, 2:31–113, 1926.

[135] Stefan Washietl, Rainer Machné, and Nick Goldman. Evolutionary footprints of nucleosome positions in yeast, 2008.

[136] Deirdre Weymann, Janessa Laskin, Robyn Roscoe, Kasmintan A Schrader, Stephen Chia, Stephen Yip, Winson Y Cheung, Karen A Gelmon, Aly Karsan, Daniel J Renouf, et al. The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. *Molecular genetics & genomic medicine*, 5(3):251–260, 2017.

[137] Sharon J. Wiback and Bernhard O. Palsson. Extreme pathway analysis of human red blood cell metabolism. *Biophysical Journal*, 2002.

[138] C. R. Woese. Order in the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 1965.

[139] C R Woese. Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences of the United States of America*, 97(15):8392–8396, 2000.

[140] Natalya Yutin, Vladimir V. Kapitonov, and Eugene V. Koonin. A new family of hybrid virophages from an animal gut metagenome. *Biology Direct*, 2015.

[141] Katarzyna Zaremba-Niedzwiedzka, Eva F. Caceres, Jimmy H. Saw, DIsa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W. Seitz, Karthik Anantharaman, Piotr Starnawski, Kasper U. Kjeldsen, Matthew B. Stott, Takuro Nunoura, Jillian F. Banfield, Andreas Schramm, Brett J. Baker, Anja Spang, and Thijs J.G. Ettema. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 2017.

[142] Jinglie Zhou, Dawei Sun, Alyson Childers, Timothy R. McDermott, Yongjie Wang, and Mark R. Liles. Three Novel Virophage Genomes Discovered from Yellowstone Lake Metagenomes. *Journal of Virology*, 2015.