Bioinformatics challenges and potentialities in studying extreme environments

Claudio Angione[1,*], Pietro Lió[2,*], Sandra Pucciarelli[3,*], Basarbatu Can[4], Maxwell Conway[2], Marina Lotti[5], Habib Bukhari[6], Alessio Mancini[3], Ugur Sezerman[4], and Andrea Telatin[7]

[1] School of Computing, Teesside University, UK
[2] Computer Laboratory, University of Cambridge, UK
[3] School of Biosciences and Veterinary Medicine, University of Camerino, Italy
[4] Epigenetiks Genetik Biyoinformatik Yazilim A.S., Turkey
[5] Department of Biotechnology and Biosciences, State University of Milano Bicocca, Italy
[6] COMSATS Institute of Information Technology, Islamabad, Pakistan
[7] BMR Genomics, Italy

*These authors contributed equally to this work

Abstract. Biological systems show impressive adaptations at extreme environments. In extreme environments, directional selection pressure mechanisms acting upon mutational events often produce functional and structural innovations. Examples are the antifreeze proteins in Antarctic fish and their lack of hemoglobin, and the thermostable properties of TAQ polymerase from thermophilic organisms. During the past decade, more than 4000 organisms have been part of genome-sequencing projects. This has enabled the retrieval of information about evolutionary relationships among all living organisms, and has increased the understanding of complex phenomena, such as evolution, adaptation, and ecology. Bioinformatics tools have allowed us to perform genome annotation, cross-comparison, and to understand the metabolic potential of living organisms. In the last few years, research in bioinformatics has started to migrate from the analysis of genomic sequences and structural biology problems to the analysis of genotype-phenotype mapping. We believe that the analysis of multi-omic information, particularly metabolic and transcriptomic data of organisms living in extreme environments, could provide important and general insights into the how natural selection in an ecosystem shapes the molecular constituents. Here we present a review of methods with the aim to bridge the gap between theoretical models, bioinformatics analysis and experimental settings. The amount of data suggests that bioinformatics could be used to investigate whether the adaptation is generated by interesting molecular inventions. We therefore review and discuss the methodology and tools to approach this challenge.

Keywords: multi-omic, multi-layer networks, adaptation, metabolism, extreme environments.

Introduction

Population genetics and multi-omics systems biology have independently witnessed increasing research attention in recent years [1, 2]. For instance, mathematical models for investigating genotype-phenotype relations have been developed for specific organisms, mainly bacteria [3, 4]. However, predictions of cellular behavior cannot disregard bioinformatics methodologies that estimate the capability of adaptation of the cell to varying environmental conditions [5]. Methods for studying the molecular response to adaptation are still lacking, and would require a multi-scale and multi-omic combination of tools commonly employed in bioinformatics, but currently used referring only to single scales or to a single omic.

This paper reviews molecular and bioinformatics methods for studying molecular adaptation. The review is divided into distinct methods/software blocks describing existing software tools and techniques (Fig. 1), further reviewed in the following sections. Note that we focus mainly on the analysis of organisms from extreme environments as they possess distinct properties that allow deciphering the basis of environmental adaptation. In fact, the evolution of genome and phenome depends on the robustness of an organism and on its ability to adapt to varying conditions [6]. For this reason, as we discuss in the following, innovations are often found in organisms living in extreme environments. The sequence of functional blocks in the figure leads to the identification of pathways, genes and proteins involved in adaptation. Each block contains distinct methodologies which could be implemented in one or more existing software tools, reviewed in the relevant sections. For the sake of clarity, each block is numbered and described below in the paper.

Throughout the paper, we will stress the need for multi-omic tools. Analogously, since many tools rely on networks to represent relations between biological entities, we will propose the use of multi-layer networks instead of single-layer networks. The strengths of our design are the following: (i) use and calibration of multi-omic and multi-layer information; (ii) use of pathway information; (iii) machine learning, bioinformatics and multi-objective optimization integrated in a powerful and novel inferential engine. The paper is structured according to the three main figures. First, we follow Fig. 1 to describe how experimental techniques produce data that need models, bioinformatics and machine learning for useful interpretations. Then, we describe the main problem of the present study, i.e. the relationship between molecular changes and selection (Fig. 2 and Fig. 3). We list below the main definitions of multi-omic terms employed in this paper:

- Multi-omics approach: Method that combines several omics data, such as genomics, proteomics and transcriptomics.
- Multi-layer network: A set of networks with a 1:1 correspondence between nodes, but different edges. This can be a useful methodology to represent relations between biological networks, e.g. from genes to proteins, and from pathways to metabolic fluxes. Networks are often a natural way to model many types of biological knowledge at different levels.

- Multi-scale model: Typically, a model that has been created by combining techniques that are valid at different scales of organization. For instance, combining a number of individual flux balance models (traditionally fine-grained single cell models) to model an entire population (which would normally be done using coarse-grained agents).

The main tools described in this paper are:

- METRADE: pipeline for building and optimizing genome-scale multi-omic models that accounts for metabolism, gene expression and codon usage at both transcriptional and translational levels. Freely available as a MATLAB toolbox [7].
- Colombos v3.0: database integrating publicly available transcriptomics data for several prokaryotic model organisms [8].
- MMETSP: database providing over 650 assembled, functionally annotated, and publicly available transcriptomes. These transcriptomes largely come from some of the more abundant and ecologically significant microbial eukaryotes in the ocean [9].
- Panoga: software used to identify the affected pathways in organisms living in extreme conditions [10]. It takes the list of significantly altered genes and their significance values and maps them to a protein-protein interaction network. Panoga is also a web-server for identification of SNP targeted pathways from genome-wide association study data. The web-server is freely available at: http://panoga.sabanciuniv.edu/.

Sampling, Next Generation Sequencing (NGS) and Ribotyping to detect microbial diversity and adaptations

One of the most important parts in studying environmental adaptation is the acquisition of data needed to perform multi-omic analysis. It is essential to carefully plan the number and location for up-taking environmental samples that will be collected, in order to avoid a number of gaps for future analysis. Furthermore, it is worth developing a protocol detecting even the least prevalent type of microorganisms from relative abundant index from different sources. Sampling biases such as selection of least turbulent sites, depth, width as well as time of the year and later on DNA extraction methods for direct isolation without culturing depending must be taken into consideration during the analysis. The analysis of metagenomes from samples collected in extreme environments, such as volcanoes, glaciers, or deep ocean waters (Fig. 1, panel 1) represents a valuable resource to study the molecular mechanisms underlying environmental adaptation. An extreme environment can also be found in the human body, e.g. gut microbiota during dietary extremes and exercise, or during metabolic diseases, for which models are available [11].

According to the extreme environment under consideration, different molecular, physiological and phenotypic strategies can be unraveled by applying multi-omic approaches. For example, in [12], a comprehensive survey of the distribution of bacteria from 213 samples was generated

from 60 stations along the horizontal and vertical salinity gradients of the Baltic Sea. This represented the first detailed taxonomic study of an indigenous brackish water microbiome composed by a diverse combination of freshwater and marine clades that appears to have adapted to the brackish conditions. Furthermore, by applying whole-genome shotgun sequencing to microbial populations collected *en masse* from the Sargasso Sea near Bermuda, it was possible to discover 148 previously unknown bacterial phylotypes and to identify over 1.2 million previously unknown genes, suggesting substantial oceanic microbial diversity [13]. Microbial community profiling is also benefiting from advanced Bayesian techniques that have proven efficient strategies when multiple species are present in the mixture sampled [14].

Different populations within the same species may adapt differently to specific environmental conditions. These ecotypes or ecospecies are usually genetically distinct geographic subspecies of organisms that typically exhibit different phenotypes. However, microbial ecotypes cannot always be recognized by obvious phenotypic differences. In the last years, several genotypic methods usually based on the small subunit (SSU) ribosomal RNA (rRNA) analysis, or the rRNA internal transcribed spacer (ITS) regions (ribotyping) have greatly enhanced our capacity to quickly identify microbial species and sometimes populations from environmental samples. Also, they can be used to compare the distribution of various microorganisms isolated from animals, humans and food.

The analysis of SSU/ITS RNA sequences is also a powerful tool to characterize symbiote/host association and to identify whether a species is widespread in its distribution, or has dispersed through recent human-mediated events. For example, SSU/ITS RNA sequences were used to assess that a ciliate species of *Stentor* genus was introduced in the Lake Garda by anthropogenic activities [15]. Recently, SSU RNA phylogenetic analysis was used to characterize a bacterial consortium associated to *Euplotes focardii*, a strictly psychrophilic marine bacteria isolated from Terra Nova Bay, in Antarctica [16]. This study indicates that the consortium is also represented by Antarctic bacteria that were probably acquired by *E. focardii* after the colonization of the Antarctic marine habitat and may have contributed to its adaptation to the extreme conditions of this environment.

Extreme environments played a key role in shaping processes that are currently used in molecular biology. At extreme temperatures, it is in fact more difficult to keep a stable DNA replication process; this explains why innovations are often found in organisms living in extreme conditions. For instance, *Taq* polymerase, which is frequently a key step for the polymerase chain reaction, originates from *Thermus aquaticus*, a thermophilic microorganism. The high discriminatory power and reproducibility of polymerase chain reaction ribotyping (PCR RT) is also used for studying outbreaks at a local level, like in healthcare centers. Nosocomial infections are one of the leading causes of death among hospitalized patients and remain a major problem in all hospitals across the world. Many types of microorganisms cause infections in humans. Therefore, understanding the microbial diversity is a fundamental goal in healthcare. An increase in incidence of a PCR RT in hospitals could

provide useful data for monitoring changes in type prevalence rates and control outbreaks [17]. In this survey, 14 *Clostridium difficile* have been isolated from nosocomial patients. PCR RT was used to prove if these microorganisms were identical. Result showed that among the isolations there was a predominant *C. difficile* lineage spreading in the hospital, with 5 out 14 identical ribotyping patterns. PCR RT has rapidly become the most widely used, straightforward and affordable typing method to detect diversities among the same microorganism species [18].

Omic datasets to measure cellular response to varying environments

Due to reduced costs and improved technology, data collection has witnessed a massive growth in speed and efficiency. For instance, multi-omic datasets can be used in association with multi-omic models to further extend, optimize and refine them [19], as well as to give insights into mechanisms of adaptation to different environmental conditions (Fig. 1). By combining network inference algorithms and experimental data derived from 445 *Escherichia coli* microarrays, Faith et al. [20] identified 1079 regulatory interactions, 741 of which were new regulators of amino acid biosynthesis, flagella biosynthesis, osmotic stress response, antibiotic resistance, and iron regulation. This approach contributed to the understanding on how organisms can adapt to changing environments.
A more comprehensive dataset of gene expression levels measured in various environmental conditions, named Colombos v3.0, has been recently published by Meysman and colleagues [8]. Colombos includes *E. coli* microarray profiles for over 4000 conditions, measured using microarrays (Affymetrix *E. coli* Genome 2.0) with raw hybridization of intensities, and RNA-seq (Illumina MiSeq) with short read sequences. The expression profiles, measured on different platforms, have been then homogenized, and the conditions have been fully annotated.
RNA-seq and microarray techniques have revolutionized gene expression studies and allowed large-scale parallel measurement of whole genome expression. Both approaches represent valuable tools for the identifications of genes that are up- or down- regulated to respond to extreme conditions. High-resolution RNA-Seq transcriptome analysis of *Deinococcus gobiensis* following UV irradiation indicated the induction of genes involved in photoreactivation and recombinational repair, together with a subset of previously uncharacterized genes [21]. The investigation of the unknown genes and pathways required for the extreme resistance phenotype will highlight the exceptional ability of *D. gobiensis* to withstand environmental harsh conditions [21], providing the groundwork for the understanding of the general mechanisms of adaptation to extreme environments.
The Marine Microbial Eukaryotic Transcriptome Sequencing Project MMETSP[1] provided over 650 assembled, functionally annotated, and publicly available transcriptomes. These transcriptomes largely come from some of

[1] http://marinemicroeukaryotes.org/

the more abundant and ecologically significant microbial eukaryotes in the ocean, and allowed the creation of a valuable benchmark against which environmental data can be analyzed [9]. By exploiting MMETSP datasets, researchers are allowed to study the evolutionary relationships among marine microbial eukaryotic clades, such as ciliates [22], and within the overall eukaryotic tree of life (Keeling et al., 2014). Furthermore, the interpretation of metatranscriptomic data generated from marine ecosystems allows us to explore the physiology and adaptation of diverse microbial eukaryotes from marine ecosystems [9].
Microarray transcriptional profiling of Arctic *Mesorhizobium* strain N33 allowed the identification of the most prominent up- and down- regulated genes under eight different temperature conditions, including both sustained and transient cold treatments, compared with cells grown at room temperature [23]. Up-regulated genes encode proteins involved in metabolite transport, transcription regulation, protein turnover, oxidoreductase activity, cryoprotection (mannitol, polyamines), fatty acid metabolism, and membrane fluidity [23]. Some genes were significantly down-regulated and classified in secretion, energy production and conversion, amino acid transport, cell motility, cell envelope and outer membrane biogenesis functions. This transcriptional profiling suggests that one of the strategy to survive under cold stress conditions is to adjust cellular function and save energy by reducing or ceasing cell growth rate.

Multi-omic models can predict cellular activity

Several computational algorithms have been developed to analyze genes and gene sets in a multi-omic fashion [24]. The main goals are detecting dependencies among genes over different conditions and unraveling gene expression programs controlled by the dynamic interactions of hundreds of transcriptional regulators [20] (Fig. 1, panel 3).
The dataset by Faith et al. [20] was mapped to a multidimensional objective space through METRADE [7], a comprehensive tool for multi-omic flux balance analysis (Fig. 1, panel 4). The Colombos dataset has been exploited to predict growth rates and secretion of chemicals of interest (acetate, formate, succinate and ethanol) by mapping each environmental condition onto a multi-omic *E. coli* model that includes underground metabolism [25]. More specifically, using a multi-level linear program and a multi-omic extension of flux-balance analysis (FBA), gene expression was mapped onto a model of *Escherichia coli*. As a result, condition-specific models of *E. coli* were generated, and their predicted growth rate and production of byproduct were assessed.
A hybrid method combining multi-omic FBA and Bayesian inference was recently proposed [26] with the aim of investigating the cellular activities of a bacterium from the transcriptomic, fluxomic and pathway standpoints under different environmental conditions. The authors integrate an augmented FBA model of *E. coli* and a Bayesian factor model to regard pathways as latent factors between environmental conditions and reaction rates. Then, they determine the degree of metabolic pathway responsiveness and detect pathway cross-correlations. They also infer

pathway activation profiles as a response to a set of environmental conditions. Finally, they use time series of gene expression profiles combined with their hybrid model in order to investigate how metabolic pathway responsiveness vary over time.

In two research works, Taffi and colleagues proposed a computational framework that integrates bioaccumulation information at the ecosystem level with genome-scale metabolic models of PCB degrading bacteria [27, 28]. The authors applied their methods to the case study of the polychlorinated biphenyls (PCBs) bioremediation in the Adriatic food web. Remarkably, they were able to discover species acting as key players in transferring pollutants in contaminated food web. In particular, the role of the bacterial strain *Pseudomonas putida* KT2440, known to be able to degrade organic compound, in the reduction of PBCs in the trophic network, was assessed in different scenarios. Interestingly, one aspect of their analysis involved a scenario computed by using a synthetic strain of *Pseudomonas* performing additional aerobic degradation pathways. Combining these computational tools allows designing effective remediation strategies for contaminated environments, which can present challenges of natural selection, and provides at the same time insights into the ecological role of microbial communities within food webs.

Genome-scale modeling and community detection of extreme environmental conditions

Omics technologies facilitate the study of organisms living at extreme conditions from different perspectives. They provide insights at genomic level, transcription level, protein level, and metabolites level. When compared to omics data from the organisms living at normal conditions, these may help understand mechanism of adaptation to extreme conditions. However, each dataset represents one portion of the whole picture and to understand the whole mechanism it is vital to integrate the data and reveal the mechanisms supported by diverse range of omics data. One way to integrate omics data is through identification of the pathways affected by each data source, and through combining the significance of affected pathways via Fisher's z-score. Panoga [10] is one of the methods that is used to identify the affected pathways in organisms living in extreme conditions. It takes the list of significantly altered genes and their significance values and maps them to a protein-protein interaction network. Then it searches for active subnetworks containing most of the affected genes. Affected KEGG pathways from the set of genes in the active subnetworks are determined and assigned significance values based on hypergeometric distribution. Combination of significance values of all the affected pathways for each type of omics data reveals affected pathways by all the data available.

In the past decade, genome-scale metabolic modeling has been successfully applied also for studying large-scale metabolic networks in microbes, with the aim of guiding rational engineering of biological systems, with applications in industrial and medical biotechnology, including antibiotic

resistance [29]. Even though antibiotics remain an essential tool for treating animal and human diseases in the 21st century, antibiotic resistance among bacterial pathogens has garnered global interest in limiting their use, and to provide actionable strategies to search and support development of alternative antimicrobial substances [30]. It is interesting to note that bacterial strains such as *Arthrobacter* and *Gillisia* sp. CAL575, producing an array of molecules with potential antimicrobial activity vs human pathogenic *Burkolderia cepacia* complex strains were isolated from Antarctica [31]. These strains represent useful models to unravel metabolic pathways responsible for the production of bioactive primary and/or secondary metabolites [32].

Using methods for community detection in networks (Fig. 1, panels 5-6), environmental conditions can be grouped according to their predicted response, which is measured in the metabolic network using multi-omic models. Interestingly, this response can be measured on different omic levels. For instance they can be evaluated on the transcriptomic and fluxomic levels, each of which can constitute a layer of a multi-layer network. This approach would enable the study of the response individually on each omic layer, but also globally, e.g. by using a network fusion approach, where layers can be weighted and the multi-layer network can be fused to a single-layer network. Finally, although not covered here, another important approach to study metabolic networks is stochastic simulation based on an approach pioneered by Gillespie [33], and relying on molecular counts to simulate the evolution of populations of chemical species [34].

Protein homology modeling and directed evolution

The proteome forms the primary link between the genome and the metabolome. As such, understanding protein function is extremely important to taking a truly multi-omic view where we understand the causal interactions between layers, rather than just finding correlations. Computational models allow us to predict protein folding, and hence functionality, and are particularly useful when combined with directed evolution, which can allow us to explore entirely new structures and their properties.

Computational methods such as homology modeling and molecular dynamic simulation can be employed for protein engineering and design. Directed evolution of enzymes and/or bacterial strains can be exploited for industrial processes [35]. For instance, protein modeling and molecular dynamic simulation can be applied to molecules from psychrophilic organisms to unravel the molecular mechanisms responsible for cold-adaptation. In [36], a computational structural analysis based on molecular dynamics (MD) was performed for three β-tubulin isotypes from the Antarctic psychrophilic ciliate *Euplotes focardii*. Tubulin eterodimers (the building block of microtubules composed of α-tubulin and β-tubulin) from psychrophilic eukaryotes can polymerize into microtubules at 4 °C, a temperature at which microtubules from mesophiles disassemble. The

structural analysis based on MD indicated that all isotypes from *E. focardii*, with respect to those of mesophilic organisms, display different flexibility properties in the regions involved in the formation of longitudinal and lateral contacts during microtubule polymerization. A higher flexibility of these regions may facilitate the formation of lateral and longitudinal contacts among heterodimers for the formation of microtubules in an energetically unfavorable environment. Given that the protein structure could be thought of as a unit of phenotype, homology modeling analysis plays a major role in the generation of testable hypothesis on selection processes (disruptive, directional, stabilizing; see Fig. 2) acting at the level of genomic coding regions. One of the most important parameters influencing selection processes is the temperature. Molecular dynamics studies provide important insights into the mechanism of activity and stability of enzymes working in extreme conditions. If the three dimensional structure of the enzyme is known, one can conduct molecular dynamics runs at variable temperatures and compare the root mean square deviation (RMSD), root mean square fluctuation (RMSF) and the radius of gyration values of the enzyme at room temperatures, elevated temperatures and at low temperatures. This approach would allow tracing the unfolding mechanism and the flexibility of the enzyme. When compared with enzymes that are working at room temperatures, these runs would reveal crucial factors for activity and stability at extreme conditions. One such study that used a MD-based approach to study the mechanism of temperature stability of thermophilic lipases, showed the importance of tryptophans involved in dimer formation and enhancement of aggregation tendency [37]. Another study conducted on the same family of lipases also revealed the importance of these tryptophans for coordination of zinc ions and the dependence of the thermostability to zinc concentrations [38]

Bioinformatics and mathematical methods play a crucial role in the experimental design and in understanding the pathways of the natural and artificial evolution of protein properties. Focused databases and computational tools to study and design evolution pathways have been developed. A promising collaboration between computational methods and evolutionary mutagenesis is envisaged in the field of de novo protein design, in which folds and functions not yet existing in nature are simulated computationally. Coupling computational design with directed evolution can help improve the performance of new proteins, as shown by designing and evolving proteins able to catalyze reactions not accessible to natural enzymes [39]. Interestingly, concepts originally developed in protein design studies such as protein folding funnels and fitness adaptive landscapes [40–42] could be further extended to multi-omic information.

Multi-omic adaptive landscapes

One of the biggest challenges of multi-omic bioinformatics is the estimation of mutual relationships between different omics. As shown in Fig. 3, one omic level Z may show a different structure over time. This can depend on a previous structure at the same level, but also on the structure

of the interacting levels X and Y. By hypothesizing a linear relation between two omic levels, and a perturbation term P affecting X, in a discrete time domain we can define the interdependencies between omic levels as

$$\mathbf{x}(t+T) = \mathbf{\Lambda}\mathbf{x}(t) + \mathbf{p}(t), \tag{1}$$

where $\mathbf{x} = (X, Y, Z)^\intercal$, $\mathbf{\Lambda} = (\lambda_{\mathbf{ij}})_{\mathbf{i,j=1,2,3}}$, $\mathbf{p} = (P, 0, 0)^\intercal$. Note that, due to the large availability of genotype/phenotype data, the parameters $(\lambda_{ij})_{i,j=1,2,3}$ can be calculated with parameter estimation techniques (e.g. minimization of root mean square error).

From a reverse engineering point of view, reconstructing systems from multi-omic data will require identifying correlations between selective micro-level identifiers and macro-level properties (for example the physiological level). To achieve this aim, we need to model structure and dynamics of the funnel to figure out requisite dimensionality (danger of losing dimensions instead of merely losing detail) and covariates of the problem space, given that many possible paths may have led to the same observed outcome. The funnel reconstruction is affected by the uncertainty due to multi-omic data analytics from different sources (quality, and conditions), i.e. the identifying the structure and dimensionality of metadata.

Although the different omics are subtly coupled, one meaningful and useful perspective from the operational standpoint, is the concept of multi-layer networks. Network theory investigates the global topology and structural patterns of the interactions among the constituents of a multi-omic adaptive system. Networks are the most natural way to model many types of biological knowledge at different levels. Recently, complex network theory has been extended towards multiple networks. Multilayered networks can be used for the representation and quantification of the interactions arising from the combined action of omics in response to mutational events and selection pressures. The architecture of complex networks is a natural embedding for fitness-changes diffusion processes as a function of natural selection constraints, e.g. those observed during environmental changes.

Conclusion

In this paper, we reviewed available methods to study selection and adaptation processes to extreme environments by means of multi-omic experimental and bioinformatics approaches. To date, a great deal of biological information has already been acquired through application of individual 'omics' approaches. However, "multi-omic technology", coupled with or mapped to multi-layer networks, will enable the integration of knowledge at different levels: from genes to proteins, from pathways to metabolic fluxes. We have also discussed how the investigation of the correlation between changes at the level of omics and the selection processes may provide a useful approach to study the genotype-phenotype mapping.

We argue that extreme conditions could provide better insights into genotype-phenotype mapping, because extreme phenotypes are required

to adapt to extreme environments. Large adaptations such as the production of vast amounts of antifreeze protein are much easier to detect and understand than the far more subtle changes one expects in more hospitable environments, where adaptation is primarily concerned with slight changes to an already near-optimal phenotype.
Multi-omic and multi-layer models represent a novel and powerful tool to generate discrete and testable biological hypotheses. We also argue that the study of life in extreme environments will provide useful clues to general laws of biological adaptations and easier conditions to test mechanistic hypotheses.

Acknowledgements

This research work was supported by the European Commission Marie Sklodowska-Curie Actions H2020 RISE Metable - 645693

References

1. Rasmus Nielsen and Montgomery Slatkin. *An introduction to population genetics: theory and applications*. Sinauer Associates Sunderland, MA, 2013.
2. James T Yurkovich and Bernhard O Palsson. Solving puzzles with missing pieces: The power of systems biology [point of view]. *Proceedings of the IEEE*, 104(1):2–7, 2016.
3. Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.
4. Javier Carrera, Raissa Estrela, Jing Luo, Navneet Rai, Athanasios Tsoukalas, and Ilias Tagkopoulos. An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of escherichia coli. *Molecular systems biology*, 10(7):735, 2014.
5. Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Rev. Genet.*, 13(7):505–516, 2012.
6. Eugene V Koonin and Yuri I Wolf. Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, 11(7):487–498, 2010.
7. C Angione and P Lió. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific Reports*, 5:15147, 2015.
8. Pieter Meysman, Paolo Sonego, Luca Bianco, Qiang Fu, Daniela Ledezma-Tejeida, Socorro Gama-Castro, Veerle Liebens, Jan Michiels, Kris Laukens, Kathleen Marchal, et al. Colombos v2. 0: an ever expanding collection of bacterial expression compendia. *Nucleic acids research*, 42(D1):D649–D653, 2014.

9. Patrick J Keeling, Fabien Burki, Heather M Wilcox, Bassem Allam, Eric E Allen, Linda A Amaral-Zettler, E Virginia Armbrust, John M Archibald, Arvind K Bharti, Callum J Bell, et al. The marine microbial eukaryote transcriptome sequencing project (mmetsp): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*, 12(6):e1001889, 2014.
10. Burcu Bakir-Gungor, Ece Egemen, and Osman Ugur Sezerman. Panoga: a web server for identification of snp-targeted pathways from genome-wide association study data. *Bioinformatics*, page btt743, 2014.
11. Fredrik Karlsson, Valentina Tremaroli, Jens Nielsen, and Fredrik Bäckhed. Assessing the human gut microbiota in metabolic diseases. *Diabetes*, 62(10):3341–3349, 2013.
12. Daniel PR Herlemann, Matthias Labrenz, Klaus Jürgens, Stefan Bertilsson, Joanna J Waniek, and Anders F Andersson. Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *The ISME journal*, 5(10):1571–1579, 2011.
13. J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, et al. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, 2004.
14. Sofia Morfopoulou and Vincent Plagnol. Bayesian mixture analysis for metagenomic community profiling. *Bioinformatics*, page btv317, 2015.
15. Sandra Pucciarelli, Federico Buonanno, Giovanna Pellegrini, Sabrina Pozzi, Patrizia Ballarini, and Cristina Miceli. Biomonitoring of lake garda: Identification of ciliate species and symbiotic algae responsible for the black-spot bloom during the summer of 2004. *Environmental Research*, 107(2):194–200, 2008.
16. Sandra Pucciarelli, Raghul Rajan Devaraj, Alessio Mancini, Patrizia Ballarini, Michele Castelli, Martina Schrallhammer, Giulio Petroni, and Cristina Miceli. Microbial consortium associated with the antarctic marine ciliate euplotes focardii: An investigation from genomic sequences. *Microbial Ecology*, pages 1–14, 2015.
17. Alessio Mancini, Daniele Verdini, Giorgio La Vigna, Claudia Recanatini, Francesca E Lombardi, and Simone Barocci. Retrospective analysis of nosocomial infections in an italian tertiary care hospital. *New Microbiologica*, 2016 (in press).
18. CW Knetsch, TD Lawley, MP Hensgens, J Corver, MW Wilcox, and EJ Kuijper. Current application and future perspectives of molecular typing methods to study clostridium difficile infections. *Euro Surveill*, 18(4):20381, 2013.
19. C Angione, J Costanza, G Carapezza, P Lió, and G Nicosia. Multi-target analysis and design of mitochondrial metabolism. *PLoS One*, 10(9):e0133825, 2015.
20. Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, 2007.

21. Menglong Yuan, Ming Chen, Wei Zhang, Wei Lu, Jin Wang, Mingkun Yang, Peng Zhao, Ran Tang, Xinna Li, Yanhua Hao, et al. Genome sequence and transcriptome analysis of the radioresistant bacterium deinococcus gobiensis: insights into the extreme environmental adaptations. *PloS one*, 7(3):e34458, 2012.
22. Yan Zhao, Zhenzhen Yi, Eleni Gentekaki, Aibin Zhan, Saleh A Al-Farraj, and Weibo Song. Utility of combining morphological characters, nuclear and. 2015.
23. Abdollah-Fardin Ghobakhlou, Anne Johnston, Linda Harris, Hani Antoun, and Serge Laberge. Microarray transcriptional profiling of arctic mesorhizobium strain n33 at low temperature provides insights into cold adaption strategies. *BMC genomics*, 16(1):383, 2015.
24. Steffen Sass, Florian Buettner, Nikola S Mueller, and Fabian J Theis. Ramona: a web application for gene set analysis on multilevel omics data. *Bioinformatics*, page btu610, 2014.
25. Claudio Angione, Max Conway, and Pietro Lió. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics*, 17(4):257, 2016.
26. Claudio Angione, Naruemon Pratanwanich, and Pietro Lió. A hybrid of metabolic flux analysis and bayesian factor modeling for multi-omics temporal pathway activation. *ACS Synthetic Biology*, page DOI:10.1021/sb5003407, 2015.
27. Marianna Taffi, Nicola Paoletti, Claudio Angione, Sandra Pucciarelli, Mauro Marini, and Pietro Liò. Bioremediation in marine ecosystems: a computational study combining ecological modeling and flux balance analysis. *Frontiers in genetics*, 5, 2014.
28. Marianna Taffi, Nicola Paoletti, Pietro Liò, Sandra Pucciarelli, and Mauro Marini. Bioaccumulation modelling and sensitivity analysis for discovering key players in contaminated food webs: the case study of PCBs in the adriatic sea. *Ecological Modelling*, 306:205–215, 2015.
29. Caroline B Milne, Pan-Jun Kim, James A Eddy, and Nathan D Price. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnology journal*, 4(12):1653–1670, 2009.
30. O Nolte. Antimicrobial resistance in the 21st century: a multifaceted challenge. *Protein and Peptide letters*, 21(4):330–335, 2014.
31. Marco Fondi, Valerio Orlandini, Isabel Maida, Elena Perrin, Maria Cristiana Papaleo, Giovanni Emiliani, Donatella De Pascale, Ermenegilda Parrilli, Maria Luisa Tutino, Luigi Michaud, et al. Draft genome sequence of the volatile organic compound-producing antarctic bacterium arthrobacter sp. strain tb23, able to inhibit cystic fibrosis pathogens belonging to the burkholderia cepacia complex. *Journal of bacteriology*, 194(22):6334–6335, 2012.
32. Valerio Orlandini, Isabel Maida, Marco Fondi, Elena Perrin, Maria Cristiana Papaleo, Emanuele Bosi, Donatella de Pascale, Maria Luisa Tutino, Luigi Michaud, Angelina Lo Giudice, et al. Genomic analysis of three sponge-associated arthrobacter antarctic strains, inhibiting the growth of burkholderia cepacia complex bacteria by synthesizing volatile organic compounds. *Microbiological research*, 169(7):593–601, 2014.

33. Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.
34. Luca Cardelli, Marta Kwiatkowska, and Luca Laurenti. Stochastic analysis of chemical reaction networks using linear noise approximation. In *Computational Methods in Systems Biology*, pages 64–76. Springer, 2015.
35. Jose L Adrio and Arnold L Demain. Microbial enzymes: Tools for biotechnological processes. *Biomolecules*, 4(1):117–139, 2014.
36. Federica Chiappori, Sandra Pucciarelli, Ivan Merelli, Patrizia Ballarini, Cristina Miceli, and Luciano Milanesi. Structural thermal adaptation of β-tubulins from the antarctic psychrophilic protozoan Euplotes focardii. *Proteins: Structure, Function, and Bioinformatics*, 80(4):1154–1166, 2012.
37. Emel Timucin and O Ugur Sezerman. Zinc modulates self-assembly of bacillus thermocatenulatus lipase. *Biochemistry*, 54(25):3901–3910, 2015.
38. Emel Timucin, Alexandra Cousido-Siah, André Mitschler, Alberto Podjarny, and Osman Ugur Sezerman. Probing the roles of two tryptophans surrounding the unique zinc coordination site in lipase family i. 5. *Proteins: Structure, Function, and Bioinformatics*, 84(1):129–142, 2016.
39. Olga Khersonsky, Daniela Röthlisberger, Andrew M Wollacott, Paul Murphy, Orly Dym, Shira Albeck, Gert Kiss, KN Houk, David Baker, and Dan S Tawfik. Optimization of the in-silico-designed kemp eliminase ke70 by computational design and directed evolution. *Journal of Molecular Biology*, 407(3):391–412, 2011.
40. María-Rocío Meini, Pablo E Tomatis, Daniel M Weinreich, and Alejandro J Vila. Quantitative description of a protein fitness landscape based on molecular features. *Molecular biology and evolution*, page msv059, 2015.
41. Stuart A. Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA, 1993.
42. Erik Svensson and Ryan Calsbeek. *The adaptive landscape in evolutionary biology*. OUP Oxford, 2012.

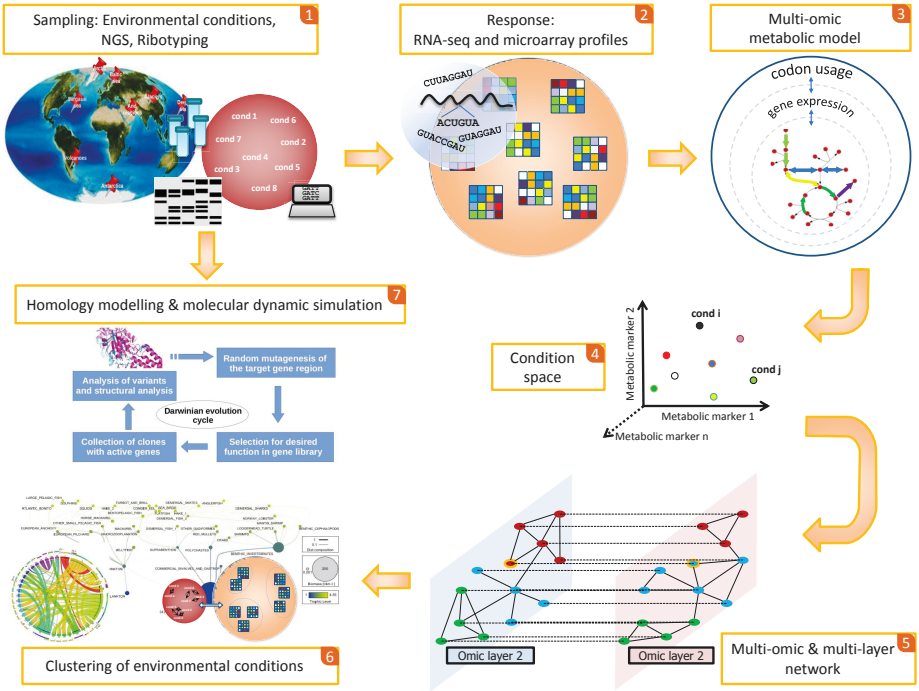

Fig. 1: **Sampling in extreme ecosystems and bioinformatics methodological applications.** The response to extreme environmental conditions (for instance Arctic and Antarctic regions, glaciers, deep ocean seawater, volcanoes and arid areas, and also similarly extreme but less exotic locations such as gut microbiota) is sampled for different associated species (1), and measured through expression profiling (2). To evaluate the environmental conditions and detect their community structure, a multi-omic model (3) can be applied to the species' metabolism, taking into account gene expression and codon usage. This model lets us map the input genotype and environment to the external behavior. The set of possible states for the organism as a response to the set of growth conditions is called *condition space* (4). Conditions can be measured for various biomarkers and therefore on various levels of omic information, e.g., a gene expression profile on the transcriptomic layer and a profile of flux rates on the fluxomic layer. Their interaction can be modeled using multi-layer networks (5). Statistical estimators and community detection methods defined on the multi-omic model can be used to investigate the pathway basis of the relationships between conditions and species in the association (6). After sampling, in parallel, homology modeling and molecular dynamic simulation can be applied to calculate structure flexibility and binding affinity of molecules of interest at different temperatures (7).

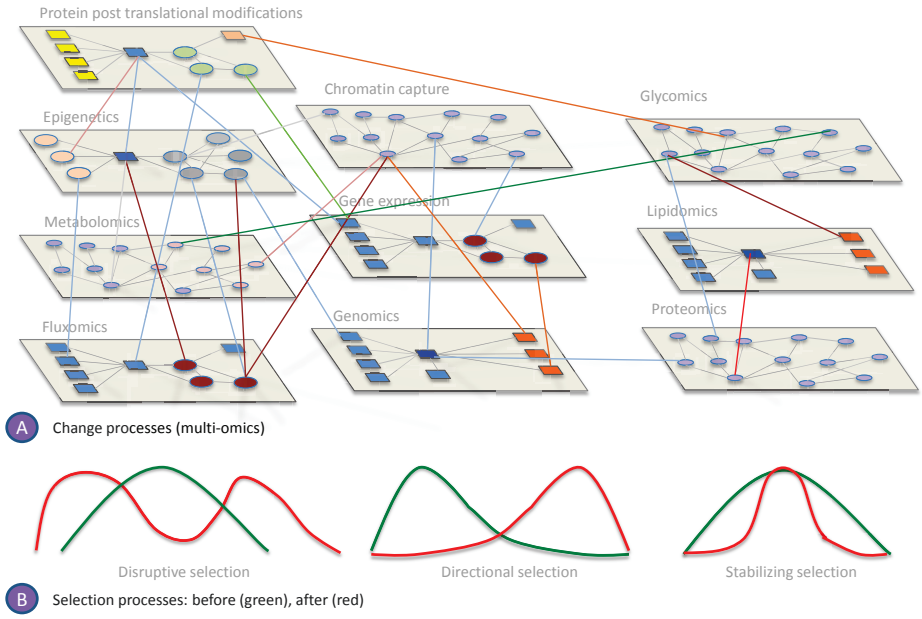

Fig. 2: **Multi-omic changes - selection problem statement.** (A) This is an extension of Fig. 1, panel 5; we show how multi-omics enlarge the molecular changes events affecting the phenotypic variations. For the sake of clarity, we show only few connections between the different omic levels. The changes are then filtered by selection processes. (B) We show the three main types of selection pressure: disruptive, directional and stabilizing. In green we show the allele distribution before the selection and in red the distribution after the selection process. It is noteworthy that with adaptation there is an increase of specificity, i.e. the number of many-to-many relationships decreases and the number of one-to-one relationships increases.

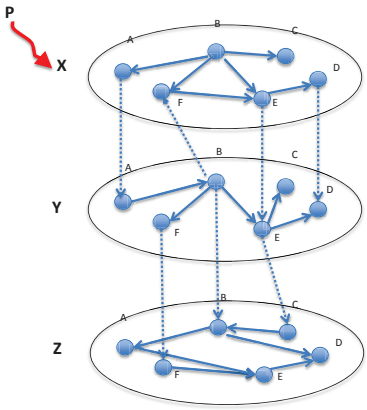

Fig. 3: **Interdependency among omic layers in a multi-layer network.** One of the challenges is to estimate the mutual relationships between omics, i.e. the causal relationships. In general terms, the control and target is an important bioinformatics challenge. Nodes represent genes forming a network, but they can also represent different features. A perturbation P may affect one or more layers. The overall response of each layer is produced by the combination of X, Y and Z, according to the interdependency among layers (see Eq. 1).