ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *Evolutionary Dynamics and Information Hierarchies in Biological Systems*

Evolutionary dynamics and information hierarchies in biological systems

Sara Imari Walker,[1,2,*] Benjamin J. Callahan,[3,*] Gaurav Arya,[4] J. David Barry,[5] Tanmoy Bhattacharya,[6,7] Sergei Grigoryev,[8] Matteo Pellegrini,[9] Karsten Rippe,[10] and Susan M. Rosenberg[11]

[1]BEYOND: Center for Fundamental Concepts in Science, Arizona State University, Tempe, Arizona. [2]Blue Marble Space Institute of Science, Seattle, Washington. [3]Department of Applied Physics, Stanford University, Stanford, California. [4]Department of NanoEngineering, University of California, San Diego, La Jolla, California. [5]Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, United Kingdom. [6]Sante Fe Institute, Sante Fe, New Mexico. [7]Grp T-2, MSB285, Los Alamos National Laboratory, Los Alamos, New Mexico. [8]Penn State University College of Medicine Department Biochemistry and Molecular Biology, Pennsylvania State University, Hershey, Pennsylvania. [9]Department of Molecular, Cell, and Developmental Biology, University of California Los Angeles, Los Angeles, California. [10]Deutsches Krebsforschungszentrum (DKFZ) and BioQuant, Research Group Genome Organization & Function, Heidelberg, Germany. [11]Departments of Molecular and Human Genetics, Biochemistry and Molecular Biology, Molecular Virology and Microbiology, and Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas

Address for correspondence: Sara Imari Walker, Ph.D. BEYOND: Center for Fundamental Concepts in Science, P.O. Box 871504, Tempe, AZ 85287. sara.i.walker@asu.edu

The study of evolution has entered a revolutionary new era, where quantitative and predictive methods are transforming the traditionally qualitative and retrospective approaches of the past. Genomic sequencing and modern computational techniques are permitting quantitative comparisons between variation in the natural world and predictions rooted in neo-Darwinian theory, revealing the shortcomings of current evolutionary theory, particularly with regard to large-scale phenomena like macroevolution. Current research spanning and uniting diverse fields and exploring the physical and chemical nature of organisms across temporal, spatial, and organizational scales is replacing the model of evolution as a passive filter selecting for random changes at the nucleotide level with a paradigm in which evolution is a dynamic process both constrained and driven by the informational architecture of organisms across scales, from DNA and chromatin regulation to interactions within and between species and the environment.

Keywords: evolution; information hierarchy; chromatin; epigenetics; viruses; networks

Introduction

The field of evolution is experiencing an exciting period, as it continues to transform from a qualitative and retrospective science into a quantitative and predictive one. Darwin's natural selection and Mendel's genetic inheritance laid the foundation for the development of population genetics and the neo-Darwinian synthesis that followed. But it is now, with the advent of modern technologies—particularly in the area of sequencing—that we are able to robustly and quantitatively compare the predictions from such theories with the variety of nature. Those quantitative comparisons make clear that large gaps exist between our current understanding of evolutionary processes and what we observe in the natural world. Much of our theory rests on highly simplified caricatures at almost every level of biological organization, from genetic to phenotypic to environmental, and perhaps unsurprisingly such theories run into greater and greater difficulty as we increase our scope. For example, we now understand quantitatively, and have confirmed

*These authors contributed equally to this work.

doi: 10.1111/nyas.12140

empirically, the evolutionary dynamics of short-term, simple adaptation in (not too large) populations. However, the grander challenges associated with what has been dubbed *macroevolution* largely remain beyond our current quantitative theories, and even short-term *microevolution* confounds us when the real world differs too much from the limited assumptions of our models (e.g., homogeneous populations with random mutations).

During August 2012, a group of physicists and biologists, with diverse backgrounds and representing a broad range of research interests, came together for a workshop on "Evolutionary Dynamics and Information Hierarchies in Biological Systems" at the Aspen Center for Physics in Aspen, Colorado to discuss these challenges and to explore new approaches. The three themed weeks of the workshop focused on the organization of DNA into chromatin, epigenetic adaptation and host/pathogen interaction, and macroevolution. Although these areas represent a wide breadth of biological phenomena, several unifying themes emerged through workshop discussions. In particular, the differences between the simplicity of our theoretical models and the complex interactions characteristic of real *physical* systems were repeatedly highlighted. Workshop discussions therefore pointed to key areas where theory and observations should aim to converge as we refine our understanding of evolution.

Among the starkest contrasts between theory and reality emerged through dialogue on the difference between the physical DNA molecule and the disembodied string of letters by which we represent it. The physical structure of DNA, how it is packed and twisted and altered, is not captured by a simple string of letters, but is vital to its function. DNA in eukaryotes typically exists in the form of chromatin, a condensed network of DNA and protein, the structure of which influences the interpretation of the string of letters. Chromatin sits in a mesoscopic regime that controls the flow of information between the microscopic nucleotides and the macroscopic phenotypic traits of the cell. And the physical and chemical heterogeneity of DNA begets heterogeneity in the mutation process, making it possible for evolution to act on the mutational process even as it is fueled by it.

A second recurring contrast was apparent when comparing the well-mixed populations, usually of fixed size, that dominate our models, and the heterogeneous population regimes of nature. This difference was vividly illustrated in discussions about host–pathogen dynamics where large intra-host populations compete against aggressive immune systems, but where the broader dynamics across hosts depends on relatively low numbers and low rates of dispersal between hosts. The interplay between these two levels of population structure leads to richer and more complex pathogen dynamics than is permitted by traditional models. Further discussions centered on the influence of population structure: even when a single population regime is appropriate, the physical and network structure of real populations can dramatically change evolutionary outcomes.

The final emerging contrast at the workshop, and perhaps the most profound, was the fundamental disconnect between our flat evolutionary models that typically focus on one biological length and/or time scale, and the multi-scale hierarchies characteristic of living organisms. For example, life organizes information in a complex hierarchy ranging from DNA sequences and chromatin regulation to cellular signaling and tissue/organ organization, and to the interactions between organisms and species. All levels of this hierarchy influence fitness, and therefore selection acts simultaneously on a variety of scales ranging from the microscale (e.g., DNA) to the macroscale (e.g., ecosystems). Evolution does not act at each scale in isolation. Workshop discussions therefore focused on connecting diverse aspects of this informational hierarchy in biological systems, and how the connections between multiple temporal and spatial scales interplay with the evolutionary process. Key questions that repeatedly arose from these discussions were: What is the structure of the biological networks that transforms chemistry (e.g., genomes, protein networks) into living organisms? And how might that structure constrain or facilitate evolution? As discussed at the workshop, answers to these questions may have implications for our understanding of the emergence of life.

These workshop themes point to the importance of novel cross-disciplinary approaches in bridging the gap between our simplified models of evolutionary processes and the reality of living organisms as physical and chemical entities. The chromatin structure of DNA requires us to go beyond sequencing to characterize it, and beyond the sequence to represent it. The evolution of pathogens requires us to

account for both their population dynamics within a host and their epidemiology across a population of hosts. A deep understanding of life cannot end with a catalogue of variation, but must also describe the framework in which variation emerges and is processed. Workshop discussions repeatedly highlighted how accounting for the physical and chemical nature of organisms at various temporal, spatial, and organizational scales can lead to new perspectives on evolution. The resulting paradigm shifts may substantially differ from the picture of evolution as a passive selective filter acting on random variation provided by the neo-Darwinian synthesis. While the workshop overview provided here cannot be an exhaustive review of such a broad range of topics, it highlights these emerging themes and the open challenges that arose through workshop discussions, and references the individual contributions of workshop participants for more in-depth discussion.

Randomness and evolvability

Several workshop discussions focused on the need for an expanded picture of evolutionary processes that goes beyond the neo-Darwinian synthesis. Neo-Darwinism, synthesizing Darwinian evolution through natural selection with Mendelian genetics, asserts that mutation is a purely random event in genomic space and blind to selective environments. However, this picture is inconsistent with biochemical reality, suggesting that a deeper understanding that integrates physical and chemical insights is required. Workshop presentations on this topic explored experiments demonstrating that some mutations are not random with respect to DNA sequence, to time, or to their potential effect on survival. In particular, emphasis was placed on the role of stress-induced mutagenesis, mutational hotspots, and the feedback between an organism and its environment in natural selection. During the discussions many fascinating questions emerged about variation in adaptive potential as a function of both environment and stress, pointing to a more dynamic picture of the mutational process than that assumed by the neo-Darwinian synthesis.

Mutation lacks foresight, but it can have hindsight

The theory of evolution states that members of natural populations vary in many ways, and that selection favors inheritance of those traits most fitted to the environment. As evolution is currently taught, new variants of genes are generated by mutations that are random with respect to their probability of being adaptive. Lynn Caporale (St. John's University) presented evidence that the assumption that "all mutation is random" is not consistent with a growing body of data;[1] she also asserted that the statement that all mutation is random is not actually consistent with the theory of evolution.

The assumption that mutations must be random with respect to adaptive value arises from the argument that processes that generate mutation have "no foresight." This would be true if environmental change was random, yet many challenges recur, such as the need to combat pathogens (and the need for pathogens to avoid host defenses). A lineage that evolves an effective response to a class of challenges would be expected to survive such repeated challenge more effectively than one that responded randomly each time.

Due to sequence-dependent variations in physical chemical properties, the probabilities of distinct classes of mutation vary along the DNA sequence. Since selection acts on variation, Caporale pointed out that evolutionary theory actually predicts that selection can act to make mutations non-random with respect to their potential effect on survival. An environment that changes in non-random ways selects for non-random variation.[2] Caporale then described multiple examples of non-random mutation. For example, pathogens with non-codon length repetitive sequences (such as CAATCAATCAATCAAT) in their coat genes generate immune-defying coat protein variants at rates 1000 times that of the background mutation rate.[3]

One widely used protocol that enables efficient storage of extensive diversity is the use of fragments of genes with identifying tags recognized by other genes or gene products. This protocol effectively encodes rules for the assembly of a diverse set of genes not explicitly encoded in the genome.[2,4] Among systems discussed at the Aspen workshop in which the use of gene fragments and rules for their assembly has evolved to encode diversity are the vertebrate immune system and trypanosome coat proteins (see e.g., the section highlighting Dave Barry's work below).

DNA sequence variation can be regulated biochemically in many ways: through altering nucleotide pools, decreasing mismatch repair,

changing the balance among different repair proteins, inducing novel polymerases, and releasing transposable elements. Caporale explained that in contrast to widely used statements of evolutionary theory, we should not assume that mutation is constant in time. One clear example of such temporal change is the increased mutation rate that can accompany stress (defined as a sensed maladaptation to the environment that results in the activation of a biochemical stress response) that was presented at the workshop by Susan Rosenberg.

Mutation as a stress response and the regulation of evolvability

Susan Rosenberg (Baylor College of Medicine) presented experimental evidence that mutagenesis is regulated in both *time*, through increased rates of mutagenesis during periods of stress, which generates new mutations specifically when cells or organisms are maladapted to their environments, and in genomic *space*, in which mutations are observed to cluster in genomes.

Although in unstressed *Escherichia coli* cells repair of double-strand breaks by homologous recombination is non-mutagenic and uses high-fidelity DNA polymerase III (Pol III), when stressed cells switch to a mutagenic mode of DNA break repair that uses error-prone DNA polymerase DinB. DinB participates in DNA break repair and generates mutations under control of the SOS DNA damage response and the RpoS-general/starvation stress response.[5,6] Rosenberg's data suggest that most spontaneous mutation in starved *E. coli* results from DNA double-strand break (DSB)–dependent stress-induced mutagenesis (SIM), requiring (1) DSBs and their repair by recombination, (2) activation of the SOS DNA damage response, which upregulates DinB levels, and (3) a separate stress such as starvation that activates the RpoS general stress response, which allows DinB DNA polymerase to participate in mutagenic break repair.[6] Interestingly, Rosenberg showed that artificial activation of the stress-response is sufficient to trigger SIM even in unstressed cells (i.e., stress itself is not required).

Various studies had previously suggested that genomic mutation hotspots might be related to DSBs,[5,7–10] but the results were open to multiple interpretations. By engineering DSBs at various sites in the genome, Rosenberg's team found that DSBs produce two distinct kinds of mutation hotspots that form by different mechanisms.[11] The first are strong local hotspots that are maximal within the first few kilobases (and extend to 60kb) of repaired DSBs, and which form via RecBCD-dependent exonucleolytic resection from DSBs and gap-filling synthesis. The second are weak long-distance hot zones extending up to approximately 1 Mb away from the DSB site, and that form independently of resection, probably via break-induced replication. That mutations are confined to local zones by the coupling of mutagenesis to DSB repair could be evolutionarily important in that it allows multiple simultaneous mutations within genes.

Rosenberg concluded with the identification of a large protein network required by *E. coli* to run the program of mutagenic repair of DNA breaks in response to stress.[12] The 93 genes identified as either promoting or required for stress-induced mutagenesis include 21 regulatory genes, 7 proteases/chaperones, 12 genes involved in DNA replication and repair, 20 genes that encode electron transfer functions, 8 genes involved in metabolism, 12 in cellular processing, and 12 of unknown function. More than half of the genes sense stress and transduce the stress signal that ultimately allows activation of three critical stress response regulators, which are key network hubs: the SOS DNA damage response, the RpoS general stress response, and the RpoE membrane protein stress response. The surprising conclusions of this study are that (1) many proteins are required to run a relatively simple program of mutagenic DNA repair; (2) most of the proteins function in stress signal transduction to key network hubs—the stress-response activators, which allow mutagenesis; and therefore (3) the large number of proteins allocated to sensing and communicating stress makes it clear that increasing mutagenesis at times of stress is important. Understanding how individual genes within this network affect stress-induced mutation is shedding important light on how SIM affects system-level evolution of protein networks, thus providing important insights into the hierarchy of control of evolvability at a systems level.

Epigenetics and chromatin organization

The discussions on the evolvability of mutation rates (discussed above) highlighted the importance of the physical and chemical structure of DNA in

understanding evolutionary processes. In eukaryotes, nucleic DNA is organized into chromatin—a highly condensed complex of DNA and protein. The structure of chromatin helps determine which stretches of DNA are read, and when, and can also influence fundamental evolutionary processes such as mutation rates.[13] However, much work remains to understand the physical structure of chromatin, especially *in vivo*, and how it is epigenetically regulated. The multiple spatial and temporal scales involved in the hierarchical processes by which the eukaryotic genome is packaged into chromatin complicate these efforts. Several workshop presentations explored current research on chromatin structure and the way in which epigenetic networks control access to chromatin. The results presented illuminate the constraints that epigenetic regulation and the physical structure of chromatin impose on evolution, thus extending our understanding of how information processing within the eukaryotic cell influences adaptive processes.

Modeling DNA organization in nucleosomes, chromatin, and chromosomes

Gaurav Arya (University of California, San Diego) presented his group's recent efforts in developing computational models that can describe the packing of DNA into chromatin at the multiple length scales at which this structuring happens. Arya began by reviewing the fascinating hierarchical process by which the eukaryotic genome is packaged. He then stressed the need to understand the organization within each hierarchical level due to the critical role of chromatin organization in modulating DNA accessibility, protein binding, and long-range genomic interactions, as well as its more apparent role in genome packaging. He particularly emphasized the importance of coarse-grained (CG) models for developing such an understanding, i.e., models capable of probing the large length and time scales of each organizational level that cannot be probed by atomistic models.

At the lowest level, eukaryotic chromatin is composed of repeats of structurally uniform units called nucleosomes, which consist of DNA segments of approximately 147 base pairs wrapped around an octamer of histone proteins. Arya developed a CG model that elucidates the dynamics of force-induced unraveling of nucleosomes, a problem relevant to DNA accessibility and nucleosome remodeling.[14] In this model, the DNA and histone octamer are treated as separate entities capable of assembling and disassembling. DNA–histone interactions are parameterized to reproduce the DNA–histone interaction free energy profile and unwrapping forces obtained from single-molecule experiments. Arya used Brownian dynamics simulations of the nucleosome where the flanking DNA were pulled apart at fixed speeds to explore the dynamics of the wrapped DNA and the motions of the histone octamer accompanying nucleosome unraveling. An important finding is the role that non-uniform DNA–histone interactions along wrapped DNA play in stabilizing nucleosomes against unraveling, while enhancing the accessibility of the wound DNA via breathing motions.

Moving beyond individual nucleosomes, Arya, in collaboration with Tamar Schlick, developed a CG model of nucleosome arrays—segments of bound nucleosomes connected by a contiguous strand of DNA—that describes their conformation and interactions using a few important degrees of freedom, while accounting for key physical features including thermal fluctuations, configurational entropy, DNA mechanics, nucleosome shape, linker histone binding, histone tail flexibility, excluded volume, and salt-screened electrostatic interactions.[15,16] Arya's Monte Carlo simulations of this model demonstrate a polymorphic structure of chromatin fibers, which fits well with the crosslinking experiments of Sergei Grigoryev (discussed in the next section), and which reveal the importance of physiological salt, histone tails, and the linker histone to the stability of the compact state of chromatin at this level of packaging.

Finally, motivated by the need to understand higher-order chromatin folding, Arya described a computational approach to determine chromatin conformations from interaction frequencies (IFs) measured by chromosome conformation capture and related techniques.[17] Dynamic simulations of a restrained bead-chain model are used to estimate the IFs. These estimates are then incorporated into an adaptive algorithm that iteratively refines the strengths of the imposed restraints on the bead-chain until a match between the computed and experimental IFs is achieved. This approach has been validated against multiple simulated test systems and is currently being refined against experiments.

Higher order chromatin folding

Sergei Grigoryev (Pennsylvania State University College of Medicine) delved further into higher-order chromatin folding, focusing on the role of the nucleosome repeat length (NRL) and nucleosome–nucleosome interactions. Nucleosome arrays fold into higher-order chromatin, which controls chromatin condensation and its accessibility for transcription, recombination, and other DNA-directed biological processes. However, as highlighted above, the architecture of higher-order chromatin is still poorly understood, and many of its physical properties are unknown.

Over recent years, Grigoryev's laboratory has extensively characterized condensed chromatin in several types of terminally differentiated cells from chicken, mouse, and human. Grigoryev discussed observations that chromatin condensation in each cell type was associated with changes in NRL, the concentrations of linker histone and tissue-specific non-histone architectural proteins, as well as post-translational histone modifications. These studies have revealed significant differences between organisms and tissue types, suggesting that the process of developmental regulation of chromatin condensation, although fundamental for cell differentiation, is not evolutionary conserved. Experiments also suggest that in some cell types, chromatin condensation may involve massive interdigitation between folded nucleosome arrays promoted by chromatin bridging factors and histone modifications that mediate nucleosome interactions between the arrays.

Grigoryev next focused on the contribution of the NRL to chromatin higher-order structure, as revealed by work on synthetic DNA sequences which vary only in the spacing between repeated nucleosome-specific DNA sub-sequences (i.e., their NRL).[18] High-resolution nuclease mapping of these sequences showed that nucleosome arrays maintain protection of DNA from nuclease activity by linker histones, consistent with formation of linker DNA stems observed by electron microscopy (EM). The use of sedimentation and EM techniques revealed an overall negative correlation between NRL and chromatin folding. In the shorter NRL range of 165–177 base pairs (typical of less condensed, transcriptionally-active chromatin), Grigoryev described a strong periodic dependence of chromatin folding on small changes in NRL. This relationship suggests that the transcriptionally active yeast genome might have evolved to have precise, short NRLs (162 and 172 base pairs) supporting relatively open higher-order structure. In contrast, the longer NRLs (188 base pairs and above) typical of vertebrate chromatin do not affect chromatin folding and need additional architectural to mediate chromatin condensation.

Grigoryev also presented studies of internal nucleosome interaction within reconstituted and native chromatin using an EM-assisted nucleosome interaction capture (EMANIC) technique.[15] For native and reconstituted chromatin condensed at physiological conditions, the experiments revealed a nucleosome interaction pattern consistent with predominantly straight linkers and a two-start helical arrangement of nucleosome cores. For the most condensed chromatin in the nuclei of terminally-differentiated cells and metaphase chromosomes, Grigoryev also discussed observations of a prevalence of nucleosome interactions typical of the two-start helix, but observed interactions also included those from folded and interdigitated chromatin fibers. The findings were discussed in relation to the mechanism of nucleosome array folding mediated by dynamic short-range nucleosome interactions, which occur even in the most condensed chromatin state.

An epigenetic mechanism to silence transcription in heterochromatin

Karsten Rippe (German Cancer Research Center) described how linking epigenetic modifications of histone residues with their readout by specific protein domains is an important aspect of current theoretical models that describe epigenetic networks.[19,20] These models are frequently characterized by a combination of feedback loops to establish bistable chromatin states: for the locus under consideration, two distinct chromatin states can stably coexist for a certain set of conditions. With respect to quantitative descriptions of epigenetic networks, three fundamental questions are particularly relevant: (1) How is the separation of the genome in active and silenced chromatin states established and maintained and what are the factors that provide specificity for distinct chromatin states? (2) How is the confinement of a given chromatin state to a certain genomic locus achieved? (3) How is a given chromatin state transmitted through the

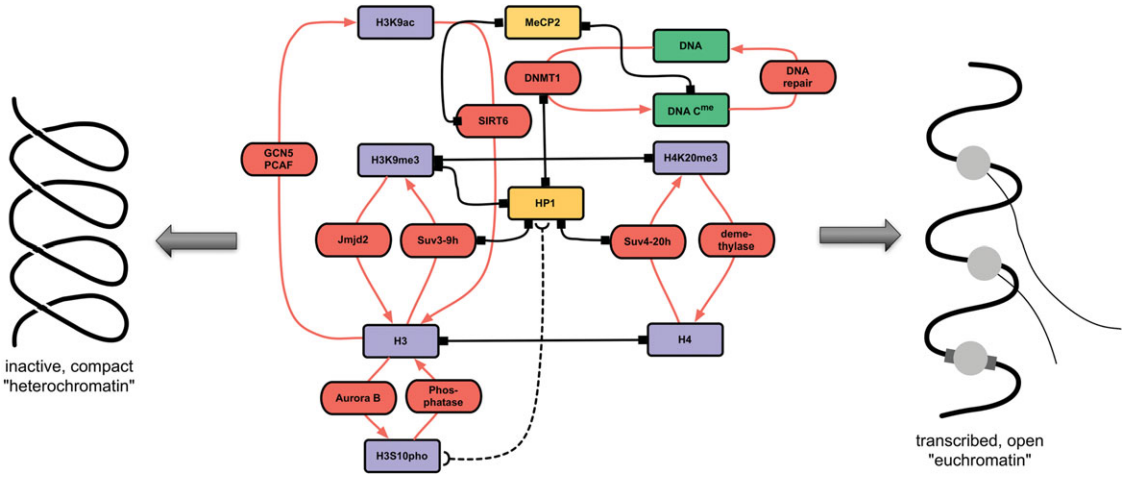

Figure 1. Regulatory epigenetic network that operates at mouse pericentric heterochromatin to silence transcription from satellite repeats. Dependent on the degree of histone H3 trimethylation at lysine 9 (H3K9me3), the system can undergo transitions between a biologically inactive heterochromatin state (high H3K9me3 levels) and an open euchromatin conformation competent for transcription (low H3K9me3 levels). The H3K9me3 modification is set by the Suv3–9h methyltransferase and is recognized by heterochromatin protein 1 (HP1). Demethylation of H3K9me3 by Jmjd2 and histone acetylation promote activation and chromatin opening. Note the linkages between the different epigenetic modifications histone trimethylation, histone acetylation and DNA methylation via protein–protein interactions of network components. Color code: red, chromatin-modifying enzymatic activities; blue, histones, with their posttranslationally methylated (me), acetylated (ac) and phosphorylated (pho) states at the indicated residues; green, DNA; yellow, structural chromatin components. The solid black lines symbolize association between proteins, while the dashed line indicates inhibition of interaction.

cell cycle, i.e., how does the cell's epigenetic memory work?

To address these questions, Rippe introduced his group's work on a particular epigenetic mechanism that silences transcription of repetitive sequences found at the pericentromeric regions (near the center of a chromosome) of the genome in mouse fibroblast cells[21] (Fig. 1). The pericentric heterochromatin (PCH) state is characterized by DNA cytosine methylation and the trimethylation of histone H3 at lysine 9 (H3K9me3) and histone H4 at lysine 20 (H4K20me3), as well as the enrichment of several chromatin modifiers and additional protein components. Studying this system to dissect epigenetic networks has the particular advantage that the PCH domains can be readily identified on fluorescence microscopy images as chromatin-dense foci. Accordingly, Rippe and colleagues were able to compare the dynamic features of PCH with the bona fide biologically active euchromatin (a state of lightly packed chromatin) that surrounds PCH by applying a previously established framework for an integrated fluorescence microscopy–based bleaching and correlation analysis in single living cells.[22,23] With this technique, they quantified protein concentrations as well as protein–chromatin and protein–protein interactions of the core protein components of PCH. The resulting comprehensive data set was used to model the epigenetic network that is active in PCH.

Rippe also discussed his group's mechanistic analysis of PCH in mouse fibroblasts as a prototypic system to explain how a repressive heterochromatin state is established and maintained. A particularly interesting result from their work is the development of a model in which the H3K9me3 modifications in PCH originate from sparsely distributed nucleation sites that distribute this modification via looping of the nucleosome chain to nucleosomes in spatial proximity. Based on previous studies, the collision probability between nucleation sites and surrounding nucleosomes was converted into a concentration[24–26] that, when cast into a model that incorporated the experimentally determined interaction parameters, yielded an excellent agreement with the measured features of mouse PCH. These results provide steps toward a quantitative understanding of epigenetic

activation and silencing and how epigenetic states are maintained.

Using high throughput sequencing to measure genome-wide chromatin structure

Matteo Pellegrini (University of California, Los Angeles) described the insights his group has gained into chromatin structure across entire genomes through next-generation sequencing technologies. A genome-wide analysis of nucleosome positions in the small flowering *Arabidopsis* plant[27] revealed that nucleosomes are enriched in certain areas of the genome, in particular in exons, and especially exon–intron boundaries. This pattern correlates with the enrichment of RNA polymerase II (Pol II) and DNA methylation in exons, consistent with a nucleosomal role in regulating Pol II processivity[28] and the targeting of DNA methylation to nucleosomes, respectively. Augmenting the nucleosome positional data with genome-wide profiles of DNA methylation demonstrated that nucleosome-bound DNA is more methylated than flanking DNA, and revealed 10-bp periodicities in the DNA methylation status of nucleosome-bound DNA. These results indicate that nucleosome positioning influences DNA methylation patterning throughout the entire genome and that DNA methyltransferases preferentially target nucleosome-bound DNA.

DNA methylation and nucleosome densities play a critical role in the regulation of gene expression,[29] but little is known about the degree to which they contribute to the differences among tissues. Pellegrini next presented tissue-specific data in which DNA methylation, nucleosome densities, and transcriptional levels were compared across tissue types. Results showed that nucleosome density is correlated with methylation and inversely correlated with gene expression. A group of root-specific genes was identified that appears to be an example of differential regulation by epigenetic marks—they are preferentially methylated, have lower nucleosome density, and at least tenfold higher expression in *Arabidopsis* roots relative to shoots—supporting a role for chromatin structure in tissue determination.

Finally, Pellegrini discussed work investigating chromatin structure on the megabase scale to identify long-range chromatin conformations, using genome-wide chromosome conformation capture (3C) coupled to high-throughput sequencing (4C-seq) to define the DNA–DNA contacts (*interactomes*) made by a number of genetic loci in pluripotent and differentiated cell types.[30] This technique enabled the identification of long-range DNA–DNA interactions that were confirmed with fluorescence *in situ* hybridization (FISH) and reciprocal 4C-seq. The data showed an organization of long-range DNA–DNA interactions specific to embryonic stem cells that is lost upon differentiation and re-established during reprogramming of differentiated cells to the induced pluripotent state. The genomic features at a given locus (i.e., transcription factor binding and chromatin state) correlate strongly with the genomic features in that locus's interactome (its entire set of molecular interactions). Pellegrini concluded by presenting evidence that chromatin could act combinatorially to guide long-range DNA–DNA interactions that account for the differences between the interactomes of pluripotent and differentiated cells, thus providing insights into epigenetic mechanisms at the level of entire genomes.

The interplay between diversity, function, and the evolution of networks

A unifying theme throughout the workshop was connecting evolutionary processes on multiple scales, ranging from the bacterial and eukaryotic genomes highlighted in the previous sections to gene networks, species and ecosystems, and even technological systems. The broad scope of these discussions encompassed a lively public panel dialog featuring workshop participants Sergei Maslov and Kim Sneppen on the topic of "Randomness and Selection in Biological and Technological Evolution", which was held the evening of August 23. The panel covered the interplay between function and randomness in examples of both biological and technological networks and stimulated much discussion among audience members including scientists and the public. Talks during the workshop also expanded on this theme. Topics covered included the roles of popularity and function in determining selection outcomes in biological (bacterial genomes) and technological (Linux installations) systems, and the evolution of networked populations in biological systems in the context of the generation and maintenance of diversity. An interesting thread connecting these diverse systems that emerged through workshop discussion is the role of network diversity and structure

(both population and spatial) in shaping evolutionary processes.

Why networks make gene families like Linux packages

Sergei Maslov (Brookhaven National Lab) explored a network-inspired analogy between the structure of biological and technological systems by comparing bacterial genomes and Linux installations. In bacteria (and Linux alike), each genome (installation) contains only a subset of the much larger universe of orthologous gene families (software packages) that are available to them by horizontal gene transfer (download). The specific subsets of the potential components observed in genomes (installations) are then the product of selection at the whole-organism (whole-computer) level for overall function, by nature (or by the user).

Maslov quantified this analogy by comparing *component frequency distributions*, defined for bacterial genomes (Linux installations) as the number of orthologs (packages) that are present at a given frequency across genomes (installations). Results from ~500 fully sequenced bacterial genomes[31] were compared with those from ~2 million Linux installations (Ubuntu Popularity Contest). In both cases the structure of the component frequency distribution can be broken into three distinct segments: (1) a large "cloud" of low frequency (present in <5% of genomes or installations) components, with numbers increasing quickly as frequency approaches zero, (2) a "shell" of intermediate frequency components, with numbers slowly decreasing as frequency increases, and (3) a small "core" of high-frequency (>90%) components found in most genomes or installations. Excluding the "core", this distribution was well fit by a power-law with an exponent of approximately –1.5, i.e., the probability (P) of an ortholog (package) being found with frequency (f) across all genomes (installations) is $p(f) \sim f^{-1.5}$.

The "cloud" and much of the "shell" consist of genes or packages that implement a variety of features, but whose function requires the presence of other genes or packages. This set of relationships can be represented in network form.[32] At the base of this dependency network are the basic and universal functions of the "core", e.g., RNA polymerase in the case of bacterial genomes or *gcc* in the case of Linux packages, with functions of increasing specificity and complexity being implemented on top of the genes or packages on which they depend (and are therefore connected to in the dependency network). Using data on the dependencies of Linux packages, Maslov showed quantitative agreement between the observed component frequency distribution and that predicted by the known dependencies among Linux packages. Maslov demonstrated that the component frequency distribution we observe across genomes (installations) is a function of the statistical properties of these dependency networks, in particular the average number of dependencies per gene (package). The results suggest that the concordance between the component frequency distributions in bacterial genomes and Linux distributions might represent a concordance in the underlying dependency networks of both bacterial genes and Linux packages, suggesting some universal organizational principles may be at work.

The origins and maintenance of diversity

Kim Sneppen (Niels Bohr Institute) discussed the role of both network and spatial structure in supporting species diversity, addressing the question: What are the minimal requirements for maintaining species diversity over long time scales? Sneppen made use of lichen ecology as a model system suitable for exploring this broad question. Lichens are organisms consisting of a symbiotic union of fungi and algae[33] that primarily inhabit surfaces. When two crustose lichen species interact on a surface, a contact boundary is formed. Diversity is maintained when these boundaries are formed between competitively equal species. Therefore, bulging boundaries observed in lichen communities suggest complex dynamics where one species may overtake another.

The model system described by Sneppen consists of a community of species competing on a two-dimensional lattice, with the ecosystem characterized by a directed network of possible species interactions.[34,35] Since species are spatially distributed, not all interactions between extant species are physically realized at a given time. The average number of species present (D) is determined by the average fraction of species that are invasible by another species (when neighbors), parameterized by γ. Sneppen introduced new species to the community at a rate α, and observed the outcome at the level of overall species diversity. The system displays a first-order phase transition at $\gamma = \gamma_c = 0.055$, transitioning from a low-diversity ($D \sim 1$) to a

high-diversity ($D \sim 20$) state as the interaction probability γ is decreased from 1 in the limit of $\alpha \to 0$.[34] Near the critical point at γ_c, the system displays bistability between low and high diversity states, where the transition between states is triggered by fragmentation of populations into patches.[35] Interestingly, these patches act as seed sites for new species to nucleate and spread.

In his discussion, Sneppen highlighted many interesting facets of the observed dynamics in this model system. One important observation was the complete collapse of population structure for systems with random neighbors (such that interactions are not determined by spatial proximity), indicating that spatial organization plays an important role in generating and maintaining diversity. Other limits discussed included introduction of long-range migration and random deaths, which both led to the loss of diversity. The results suggest a complex dynamic between network and spatial structure that dictates species diversity, and leaves many open questions. Among these, it is unclear what determines the length-scale cut-off for patch-size. From an evolutionary perspective, a particularly interesting facet of the models discussed by Sneppen is that they do not rely on a predefined fitness landscape. Instead, the fitness of an individual species is determined dynamically by its interactions with the local networked community within which it is embedded.

Coevolution and the evolutionary arms race

The co-evolutionary arms race between pathogens and the adaptive immune systems of mammals, particularly humans, was the subject of several presentations and much informal discussion at the workshop. The intense pressure applied by the mammalian immune system has driven remarkable architectural changes in the pathogens that must cope with it, as dramatically illustrated by Dave Barry in the case of *Trypanosoma brucei*—the parasitic cause of sleeping sickness. The implications of the combination of the fast and intense in-host arms race with slower dynamics at larger scales, such as the worldwide flu pandemic, and longer times, such as observed in the chronic stage of HIV, were also discussed, with a focus on current challenges in understanding and predicting the evolution of pathogens experiencing these heterogeneous regimes in space, time, and population structure.

The combinatorial diversity arsenal of the sleeping sickness parasite

Dave Barry (University of Glasgow) introduced the anti-immune system of *Trypanosoma brucei,* the single-celled protozoan that causes African sleeping sickness in humans. Trypanosomes have a dense coat of variant surface glycoprotein (VSG), which physically shields them from antibodies against conserved surface proteins and innate immunity mechanisms.[36] The only target for the immune system is VSG itself, but this is a quickly shifting target over the course of infection due to rapid switching of expression among VSG genes.[37] Trypanosomes have a very large potential for variation as VSG seems to have no function other than physically coating the cell.[38] Many VSG variants, with differences of often 80% or more at the amino acid level, appear over the course of an infection. This variation is encoded in the genome as an "archive" of thousands of VSG genes, most of which are non-functional (pseudogenes).[39] These genes sit in the mutable subtelomeric regions near the end of chromosomes, and up to 200 can be found in mini-chromosomes that contain only *VSG* genes.[40] Singular expression of VSG by each trypanosome is achieved by transcription being restricted to only a few loci, to which genes must move to become active.

Switching, which changes which VSG is expressed at a rate of $\sim 10^{-3}$ per generation, involves a process possibly initiated by a DSB at the expression site, followed by DNA repair–mediated replacement of the expressed gene with a copy of any accessible VSG variant in the genome.[41] In trypanosomes this gene conversion process often results in mosaic genes, pieced together from stretches of more than one genomic VSG variant. Thus, even pseudogenes can contribute, and the combinatorial nature of their involvement compounds with the already staggeringly large diversity that is latent within each trypanosome genome.

Archive VSG genes mutate at a high rate, using processes of gene duplication, base substitution, insertions and deletions, and segmental conversion. Their elevated mutation rate appears to be the result of second-order selection: sequence analysis shows that the subtelomeres, in which VSG genes are primarily located, mutate several-fold faster than chromosome cores. Indeed, subtelomeres are increasingly being seen as havens for diversification of eukaryotic multigene families that encode

hyper-variable phenotypes. How trypanosome cores and subtelomeres achieve fundamental mutational differences might be linked to epigenetic differences, such as base modification[41] and binding of the ORC1 replication protein.[43,44]

Despite the stochastic disorder underlying VSG switching, antigenic variation is structured. For example, variants are expressed in partially predictable order. Organization is thought to be essential for trypanosomes to deal effectively with the primary sources of selective pressure—the immune system. Modeling studies have recently begun to reveal a network of interactions from the trypanosome genome, through population dynamics, to host populations in the field. This integrative approach engenders not only predictions of how pathogen population processes are determined by underlying molecular genetics, but also inferences about resulting selective pressures on antigen gene archives.

Influenza: genomics of a "Red Queen's race"

Michael Lässig (University of Cologne) discussed the rapid evolution of seasonal influenza A virus (H3N2). Influenza evolves at a rate 10^5 times greater than *Drosophila*, and approximately 25% of nucleotides in the influenza virus have mutated since 1968. Lässig described how such rapid evolution results from immune selection that drives an evolutionary arms race between the pathogen and its host. A striking feature of this process is its punctuated pattern. Adaptive changes occur primarily in antigenic epitopes, i.e., the antibody-binding domains of the viral hemagglutinin. This process involves recurrent selective sweeps, in which clusters of simultaneous nucleotide fixations in the hemagglutinin coding sequence are observed about every four years, with a corresponding drop in diversity. The evolutionary origins of this pattern remain controversial.

Lässig suggested that the rapid adaptation of the influenza A virus produces clonal interference within the hemagglutinin gene that results in the observed punctuated pattern resulting from recurrent selective sweeps of the population.[45] Influenza A might therefore undergo a mode of evolution driven by a "Red Queen's race" between viral strains with different beneficial mutations. To infer selection under clonal interference, Lässig introduced a new method that relies on two measures: a frequency propagator ($G(x)$) defining the likelihood that a new allele reaches a frequency larger than x at some later time, and a loss propagator ($H(x)$) defining the likelihood that a new allele reaches frequencies exceeding a given threshold x at some intermediate point of its lifetime but is eventually lost. By evaluating either nonsynonymous ($G(x)$ and $H(x)$) or synonymous ($G_0(x)$ and $H_0(x)$) mutations for a sample of influenza genome sequences taken over the past 39 years, Lässig described identifying the presence of clonal interference if two conditions are fulfilled: $G(x) / G_0(x) > 1$ and $H(x) / H_0(x) > 1$ for intermediate and large frequency x. The first condition signals predominantly positive selection for a class of mutations, and the second indicates interference interactions: new beneficial alleles rise to substantial frequency, but are eventually driven to loss by a competing clone. Lässig showed that the data for influenza are consistent with clonal interference, with at least one strongly beneficial amino acid substitution per year, where a given selective sweep has on average three to four driving mutations.

In the discussion of the broader implications of this work, Lässig noted that the results strongly suggest that the course of influenza evolution is determined not only by antigenic changes, but also evolutionary competition and selection among viral strain variants: successful viral strains are those that maximize the total fitness of antigen–antibody interactions and of other viral functions by a joint process of adaptation and conservation. He also noted that calculations of the fitness flux[46] suggest that an increase in immune challenge would strongly compromise the viability of influenza. Thus, Lässig concluded that while antigenic adaptation has been a focus of influenza research so far, a broader picture of viral function and fitness is needed to understand the processes shaping influenza evolution.

Viral evolution in response to immunity

Tanmoy Bhattacharya (Sante Fe Institute and Los Alamos National Lab) discussed how viruses evolve to escape immunity, using HIV as a case study. Simian immunodeficiency viruses (SIVs), such as HIV, are lentiviruses (retroviruses with a long incubation period, capable of infecting non-dividing cells) that have coevolved with their hosts for millions of years[47] and rarely produce acute diseases in their natural hosts, but can be pathogenic in other hosts.[48] HIV jumped to humans from chimpanzees in the Congo region of Western Africa[49] around

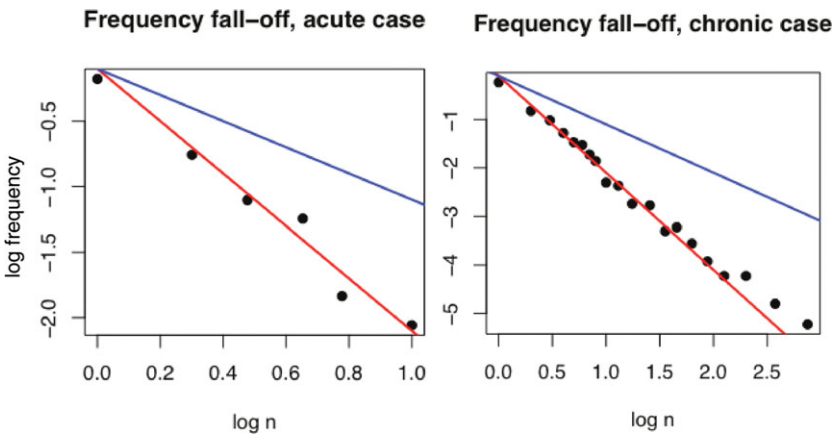

Figure 2. **Both the exponentially growing HIV population and acute infection and the constant-sized population from a chronic infection show the characteristic inverse-square fall-off of number of clones with a certain clone size. This indicates repeated rapid population replacements even during the chronic phase.**

100 years ago,[50] and is responsible for AIDS and death.[51,52] After an initial period of slow growth, the epidemic has been growing exponentially since the 1960s.[53] During this period, it broke up into roughly geographically separated subtypes, with distinct genetic signatures inherited from the clade founders.

With a mutation rate of about one change per genome every three replications,[54] a high reproductive number of about 10,[55] and a short generation time of about two days,[56] HIV is expected to adapt very quickly to its environment. Early indications that the majority of the sites in the HIV genome were under intense and consistent selective pressure from the host immune system[57] turned out to be based on an incorrect handling of the ancestral genetic differences between the subtypes.[58] In fact, the diversity of the human immune system makes only very few sites display overt adaptation signatures in population-level data.[58] Nevertheless, deep sequencing data provide indications of unexpectedly early and fast escapes from host immune reaction, often resulting in fixation of escape mutations on the time scale of a few weeks.[59]

This difference between a diversifying evolution in the population over 100 years and the "Red Queen" dynamics of the virus against the host immune system during a single untreated infection lasting about a decade has been studied in great detail at the phenotypic level.[60] Studies of clonal distributions at the sequence level also indicate that rapid selective sweeps persist during the chronic phase (Fig. 2). Furthermore, the phylogenetic trees of random samples look remarkably like the population-level influenza evolution over decades (see e.g., Fig. 7 of Lee *et al.*[60]). The implications of these fast processes for the slow adaptive journey of the HIV virus in its host are a current subject of study.

Information hierarchies, architecture, and constraints

Biological systems utilize a variety of mechanisms at various length and time scales to store, interpret, and use information. As discussed throughout this report, the information itself is organized in a complex hierarchy: from DNA sequences, to chromatin, to tissue/organ organization, to the interactions between organisms and between species. These informational hierarchies yield layered architectures that constrain evolutionary processes at multiple levels of organization. Presentations and discussions throughout the workshop explored the informational architecture of living systems and classes of functions and architectures ranging from bacterial genomes, to the human brain and the Internet, to whether identifying general principles for structuring informational hierarchies in living systems could provide insights into the emergence of life itself.

Formation of neural networks: nature versus nurture

Alexei Koulakov (Cold Spring Harbor Laboratory) addressed the roles of nature versus nurture in the development of neural systems. Koulakov opened with a discussion of the possible role of cooking in

the development of large brains in humans.[61] Brains are metabolically expensive: brain mass scales with body mass to the ¾ power (and hence also scales with metabolic rate via Kleiber's Law for metabolic scaling). This fostered a lively discussion and debate among the workshop participants about the role of cooking as a technological way of externalizing part of the digestive track. The heavy metabolic load necessary to maintain a large brain prompted Koulakov to pose the question: why have a brain?

One possible reason, as described by Koulakov, is explained by the expensive genome hypothesis, which suggests that the brain (or nervous system) is an evolutionary way to externalize parts of the adaptation process (with respect to the genome). Support for this viewpoint derives from comparing the amount of data stored in the human genome with its 3×10^9 base pairs—approximately 1 GB of data, assuming that DNA is just a linear string of bits—to the data storage capacity of the brain. Koulakov argued that the human brain, by contrast to the human genome, could store upwards of potentially 600 TB of data. This spurred the dilemma of how the approximately 1 GB of genomic information taken for a linear sequence could establish the 600 TB of information in neural network connections. Koulakov suggested that the dilemma is only resolved if synapses are not specified individually by the genome. As such, the lower-level data storage of the genome should specify a process rather than the final state of synaptic connections within the brain. He then concluded with a discussion of how models akin to those used in condensed-matter physics might be used to describe the formation of neural networks without recourse to identifying the underlying genetic mechanisms, thereby suggesting that higher-level information processing that goes beyond the flat depiction of the genome as a string of letters, as discussed throughout this report, may play an important role in the development of neural networks.

Architecture: constraints that deconstrain

John Doyle (Caltech) introduced insights into how universal laws and architecture (modularity and protocols) constrain the function and evolution of systems ranging from the biological to the technological. Several key concepts discussed by Doyle included nonconvex optimization, layered architectures as "constraints that deconstrain", and hard limits ("universal laws") on robustness and efficiency. Doyle described nonconvex optimization as a feature of robust systems given that they are large (high-dimensional) but thin (even more highly codimensional) and nonconvex in the space of all possible systems.[62] This concept is analogous to the set of words in most languages, which is both large and vanishingly thin as a fraction of all possible sequences of letters. Doyle discussed how these ubiquitous features of robust systems suggest that the path toward higher levels of organization is through protocols. Doyle explained how robust architectures are constrained by protocols, but the resulting crypticity and modularity that these constraints enable also deconstrain systems designed using such architectures, enabling them to perform more diverse tasks. What emerges from this perspective is a view of architecture as a set of constraints that enables greater flexibility by increasing the accessibility of useful alternatives.

Doyle cited an interesting consequence of "universal laws" (constraints) that emerge from layered architectures: many, like the virtualization of operating systems in modern computers, are independent of the underlying physics (e.g., hardware) and therefore result in "undecidability" rather than being predictable outcomes of the underlying physical law. Such virtualization is readily apparent in biology; however, the situation in living systems is much more complex than that for technology. The example Doyle provided was of an *E. coli* cell, which is running an incredible number of "applications" relative to more familiar operating systems such as Android or iOS. The internet was also described in juxtaposition to biology in terms of robust design. Doyle noted that a key vulnerability of the internet is that TCP/IP is not layered strongly enough, lacking modern naming and virtual addressing, thus making it vulnerable to attack.[63]

The discussion then moved to the role of hard-limits, or universal laws in robust efficiency. Doyle presented glycolysis as case study demonstrating the hard trade-offs between robustness and efficiency.[64] This led to a discussion on how much of biology relies on robust control, where most genes code for control and regulatory function. Control mechanisms are put in place to manage different resources at each layer in a given architecture, where cross-layer interactions do not occur without a programmable interface. Doyle noted that conservation

requires maintaining protocols, but once higher layers are added, flexibility allows lower levels to change. He concluded by suggesting that this might imply a conservation of change principle whereby some components of layered architectures are frozen to permit change of others. This could potentially have fascinating consequences for our understanding of evolutionary processes in biology.

The information hierarchy and the emergence of life

Sara Imari Walker (Arizona State University) discussed how insights from the organizational and hierarchical structure of biological systems provide new approaches to understanding the emergence of life. She began with some background on traditional approaches to the origin of life, including a survey of both genetics-first and metabolism-first perspectives. She then discussed how the standard criterion for judging the validity of these scenarios is the capacity to undergo Darwinian evolution, leading many to favor genetics-first perspectives under the current paradigm. She explained that while Darwinian evolution might be necessary to life, it isn't a sufficient condition for defining the transition from non-living to living systems. Walker instead highlighted the key role played by *information* in the logical reorganization of systems that make the transition to the living state. She suggested a framework for defining the origin of life as a transition in how information is managed and processed, corresponding to a transition in the causal flow of information within a physical system.[65]

As an illustrative example, Walker cited the early work of John von Neumann on self-reproducing machines and the distinction that must be drawn between trivial and non-trivial self-replication. She discussed how trivial self-replicators are systems that strictly rely on the implicit physics and chemistry of their host environment to support replication. Examples include memes, crystals, lipid composomes, and the non-enzymatic template–directed replication of nucleic acids. She contrasted these systems with non-trivial self-replicators such as living systems, which are distinct from trivial replicating systems in that they have some autonomy from their host environment. She described how this feature is characteristic of non-trivial self-replicators because they implement the active use of information via the readout of coded instructions to operate. Thus trivial and non-trivial replicators differ fundamentally in the way information is organized and how it flows through the system. Examples of non-trivial systems include a von Neumann self-reproducing machine and all known life. An important note is that both trivial (e.g., memes and non-enzymatic template replicators) and non-trivial self-replicators (e.g., living organisms) can be capable of Darwinian evolution, thus evolutionary capacity does not draw a dividing line between the two classes of systems. Walker suggested that the trivial/non-trivial distinction might provide a more rigorous criterion for defining the transition from non-life to life than the Darwinian one because it relies on identifying the presence of informational hierarchies, which may be more universally characteristic of life. Discussion among the group led to the intriguing implication that the transition from non-life to life might be undecidable in the logical sense due to the strong parallels drawn between living systems and universal computing machines.

Digging deeper into the nature of this distinction, Walker discussed how the logical and organizational structure of non-trivial replicators suggests that life might be uniquely characterized by informational architecture. In this framework, candidate measures for the transition from non-life to life would rely on the causal efficacy of distributed control, which could be characterized by top-down information flow from higher to lower levels of organization in biological informational hierarchies.[66] Walker concluded by discussing how further development of this approach might lead to novel insights and approaches into understanding the emergence of life that go beyond the specific chemical substrate of life as we now know it.

Concluding remarks

The workshop brought together leading scientists exploring the structure, function, and evolution of biological systems at various length and time scales. While perhaps not yet reaching the legal standard of a "preponderance of the evidence", the theory and experiments discussed at the workshop clearly suggest that evolution is empowered by organismal architecture at multiple scales, ranging from DNA sequences and chromatin regulation all the way to interactions between organisms and species. It is not sufficient to characterize evolution as a passive selective filter of random flips along a string

of letters: evolution is a much more dynamic process intimately intertwined with the constraints imposed by the informational architecture of biological systems and the interaction between an organism and its host environment. This view presents new challenges but also new opportunities, as discussed throughout this report. In particular, identifying how evolvability depends on the level of organization, versus being a scale-independent architectural feature, remains a challenging open question to be addressed. The challenge moving forward will be to develop new frameworks that can integrate the multiple scales of organization discussed throughout the workshop into quantitative predictions about the evolutionary process.

A highlight of the workshop was a lunchtime visit by a local baby bear at the Aspen Center for Physics. This visit was much to the delight of the workshop participants who were out enjoying their lunch in the fresh Colorado air, and to the dismay of those who missed the opportunity! However, a larger part of the excitement during the workshop was generated by the coming together of scientists from so many different disciplines, each bringing their own unique perspective on the interaction between evolution and information hierarchies in biological systems. The discussions generated by these interdisciplinary and cross-scale conversations provided fresh insights into many open questions surrounding organismal storage, use, and interpretation of information at multiple scales and the resultant impact on adaption. By sharing these discussions and open questions here, we hope that this report will further advance new approaches into this challenging subject by fostering similar cross-disciplinary discussion in the wider scientific community.

Acknowledgements

The Evolutionary Dynamics and Information Hierarchies workshop was held on August 19–September 9, 2012 at the Aspen Center for Physics in Aspen, CO. The authors wish to thank the hospitality of the Aspen Center for Physics for hosting the workshop, which was supported in part by the National Science Foundation under Grant no. PHY-1066293. The workshop was organized by Gyan Bhanot (Rutgers University), Lynn Caporale (St. John's University), Sebastian Doniach (Stanford University), Alexandre Morozov (Rutgers University), and Matteo Pellegrini (University of California, Los Angeles).

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Caporale, L.H. 2003. Natural selection and the emergence of a mutation phenotype. *Annu. Rev. Microbiol.* **57:** 467–485.
2. Caporale, L.H. 2006. An Overview of The Implicit Genome. In L. Caporale (Ed.), *The Implicit Genome.*, Oxford University Press. New York
3. Palmer, M., M. Lipsitch, E. Moxon & C. Bayliss. 2013. Broad conditions favor the evolution of phase-variable Loci. *mBio.* **4**(1): e00430–12.
4. Doyle, J., M. Csete & L. Caporale. 2006. An Engineering Perspective: The Implicit Protocols. In L. Caporale (Ed.), *The Implicit Genome.* Oxford University Press. New York.
5. Ponder, R., N. Fonville, & S. Rosenberg. 2005. A switch from high-fidelity to error-prone DNA double-strand break repair underlies stress-induced mutation. *Mol. Cell.* **19:** 791–804.
6. Shee, C., J. Gibson, M. Darrow, C. Gonzalez & S. Rosenberg. 2011. Impact of stress-inducible switch to mutagenic repair of DNA breaks on mutation in Escherichia coli. *Proc. Natl. Acad. Sci. USA.* **108:** 13659.
7. Harris, R.S., S. Longerich & S.M. Rosenberg. 1994. Recombination in adaptive mutation. *Science* **8:** 258–260.
8. Strathern, J.N., B.K. Shafer & C.B. McGill. 1995. *Genetics* **140:** 965–972.
9. Nik-Zainal, S. *et al.* 2012. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell.* **149:** 979–993.
10. Roberts, S.A. *et al.* 2012. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Mol. Cell* **46:** 424–435.
11. Shee, C., J. Gibson, & S. Rosenberg. 2012. Two Mechanisms Produce Mutation Hotspots at DNA Breaks in Escherichia Coli. *Cell Rep.* **2:** 714–721.
12. Al Mamun, A.A.M. *et al.* 2012. Identity and Function of a Large Gene Network Underlying Mutagenic Repair of DNA Breaks. *Science* **338:** 1344–1348.
13. Schuster-Bockler, B. & B. Lehner. 2012. Chromatin organization is a major influence onf regional mutation rates in human cancer cells. *Nature* **488:** 504–507.
14. Dobrovolskaia, I. & G. Arya. 2012. Dynamics of forced nucleosome unraveling and the role of nonuniform histone-DNA interactions. *Biophys. J.* **103:** 989–998.
15. Grigoryev, S., G. Arya C., Correll, *et al.* 2009. Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions. *Proc. Natl. Acad. Sci. USA* **106:** 13317–13322.
16. Arya, G. & T. Schlick. 2009. A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. *J. Phy. Chem. A* **113:** 4045–4059.
17. Meluzzi, D. & G. Arya. 2013. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Res.* **41:** 63–75.

18. Correll, S., M.H. Schubert, & S. Grigoryev. 2012. Short nucleosome repeats impose rotational modulation on chromatin fibre folding. *EMBO J.* **31:** 2416–2426.
19. Dodd, I., M. Micheelsen, K. Sneppen & G. Thon. 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell* **129:** 813–822.
20. Angel, A., J. Song, C. Dean & M. Howard. 2011. A Polycomb-based switch underlying quantitative epigenetic memory. *Nature* **476:** 105–108.
21. Probst, A.V. & G. Almouzni. 2008. Pericentric heterochromatin: dynamic organization during early development in mammals. *Differentiation* **76:** 15–23.
22. Müller, K.P. *et al.* 2009. Multiscale analysis of dynamics and interactions of heterochromatin protein 1 by fluorescence fluctuation microscopy. *Biophys. J.* **97:** 2876–2885.
23. Erdel, F., K. Müller-Ott, M. Baum, *et al.* 2011. Dissecting chromatin interactions in living cells from protein mobility maps. *Chromosome Res.* **19:** 99–115.
24. Rippe, K. 2001. Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.* **26:** 733–740.
25. Rippe, K., P.H. Hippel & J. Langowski. 1995. Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.* **20:** 500–506.
26. Ringrose, L., S. Chabanis, P.-O. Angrand, *et al.* 1999. Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. *EMBO J.* **18:** 6630–6641.
27. Chodavarapu, R.K. *et al.* 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466:** 388–392.
28. Schwartz, S., E. Meshorer & G. Ast. 2009. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16:** 990–995.
29. Berger, S.L. 2007. The complex language of chromatin regulation during transcription. *Nature* **447:** 407–412.
30. Werken, H.J. *et al.* 2012. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9:** 969–972.
31. Powell, S., D. Szklarczyk, K. Trachana, *et al.* 2012. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids. Res.* **40:** D284–D289.
32. Pang, T. & S. Maslov. 2011. A Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology. *PLoS Comput. Biol.* **7:** e1001137.
33. Nash, T. 2008. *Lichen Biology.* Cambridge University Press. Cambridge.
34. Mathiesen, J., N. Mitarai, K. Sneppen, & A. Trusina. 2011. Ecosystems with Mutually Exclusive Interactions Self-Organize to a State of High Diversity. *Phys. Rev. Lett.* **107:** 188101.
35. Mitarai, N., J. Mathiesen & K. Sneppen. 2012. Emergence of diversity in a model ecosystem. *Phys. Rev. E* **86:** 011929.
36. Schwede, A. & M. Carrington. 2010. Bloodstream form Trypanosome plasma membrane proteins: antigenic variation and invariant antigens. *Parasitology* **137:** 2029–2039.
37. Barry, J. & R. McCulloch. 2001. Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite. *Adv. Parasitol.* **49:** 1–70.
38. Barry, J. D., J.P. Hall & L. Plenderleith. 2012. Genome hyperevolution and the success of a parasite. *Ann. N.Y. Acad. Sci.* **1267:** 11–17.
39. Berriman, M. *et al.* 2005. The genome of the African trypanosome Trypanosoma brucei. *Science* **309:** 416–422.
40. Marcello, L. & J. Barry. 2007. Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. *Genome Res.* **17:** 1344–1352.
41. Boothroyd, C.E. *et al.* 2009. A yeast-endonuclease-generated DNA break induces antigenic switching in Trypanosoma brucei. *Nature* **459:** 278–281.
42. Cliffe, L., T. Siegel, M. Marshall, *et al.* 2010. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycitronic transcription units throughout the genome of Typanosome brucei. *Nuc. Acids. Res.* **38:** 3923–3935.
43. Tiengwe, C. *et al.* 2012. Genome-wide analysis reveals extensive functional interaction between DNA replication initiation and transcription in the genome of Trypanosome bruce. *Cell Rep.* **2:** 185–197.
44. Tiengwe, C. *et al.* 2012. Identification of ORC1/CDC6-interacting factors in Trypanosoma brucei reveals critical features of origin recognition complex architecture. *PLoS One* **7:** e32674.
45. Strelkowa, N. & M. Lässig. 2012. Clonal Interference in the Evolution of Influenza. *Genetics* **192:** 671–682.
46. Mustonen, V. & M. Lässig. 2010. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci. USA* **107:** 4248–4253.
47. Foley, B. 2000. An overview of the molecular phylogeny of lentiviruses. In Kuiken, C. *et al.* (Ed.), *HIV Sequence Compendium 200: Theoretical Biology and Biophysics* (pp. 35–43). Los Alamos National Lab. Los Alamos.
48. Keele, B. *et al.* 2009. Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460:** 515–519.
49. Keele, B. *et al.* 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313:** 523–526.
50. Korber, B. *et al.* 2000. Timing the Ancestor of the HIV-1 Pandemic Strains. *Science* **288:** 1789–1796.
51. Barre-Sinoussi, F. *et al.* 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220:** 868–871.
52. Gallo, R. *et al.* 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* **224:** 500–502.
53. Yusim, K. *et al.* 2001. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Proc. Roy. Soc. B* **356:** 855–866.
54. Mansky, L. & H. Temin. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69**(8): 5087–5094.
55. Stafford, M., L. Corey, Y. Cao, *et al.* 2000. Modeling Plasma Virus Concentration during Primary HIV Infection. *J. Theo. Biol.* **203:** 285–301.

56. Moore, C., M. John, I. James, *et al.* 2002. Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science* **296:** 1439.
57. Bhattacharya, T. *et al.* 2007. Founder Effects in the Assessment of HIV Polymorphisms and HLA Allele Associations. *Science* **315:** 1583–1586.
58. Fischer, W. *et al.* 2010. Transmission of Single HIV-1 Genomes and Dynamics of Early Immune Escape Revealed by Ultra-Deep Sequencing. *PLoS One* **5:** e1000890.
59. Richman, D., T. Wrin, S. Little & C. Petropoulos. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl. Acad. Sci. USA* **100:** 4144–4149.
60. Lee, H., A. Perelson, S.-C. Park & T. Leitner. 2008. Dynamic Correlation between Intrahost HIV-1 Quasispecies Evolution and Disease Progression. *PLoS Compt. Biol.* **4:** e1000240.
61. Aiello, L. & P. Wheeler. 1995. The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution. *Current Anthropology* **36:** 199–221.
62. Doyle, J. & M. Csete. 2011. Architecture, constraints, and behavior. *Proc. Natl. Acad. Sci. USA* **108:** 15624.
63. Doyle, J. *et al.* 2005. The "robust yet fragile" nature of the Internet. *Proc. Natl. Acad. Sci. USA* **102:** 14497.
64. Chandra, F., G. Buzi & J. Doyle. 2011. Glycolytic oscillations and limits on robust efficiency. *Science* **333:** 187–192.
65. Walker, S.I. & P. Davies. 2013. The algorithmic origins of life. *J. Roy. Soc. Interface* **10:** 20120869.
66. Walker, S.I., L. Cisneros & P. Davies. 2012. Evolutionary transitions and top-down causation. *Proceedings of Artificial Life XIII* (pp. 283–290). MIT Press. Cambridge.