GMR The Original

Deep Learning Integration of Multi-Species Functional Genomics to Reveal Conserved Gene Regulatory Logic

Dildora Mirakbarova, Zilola Shukurova, Davlat Dilmurodov, Sobir Xamrakulov, Dilnoza Jumanazarova, Dilbar Najmutdinova, Narzikul Maxmudov,

Associate Professor, Department of Social Sciences and Education, Tashkent International University. Tashkent, Uzbekistan , ORCID: https://orcid.org/0000-0003-3279-1794 E-mail: dildoramirakbarova07@gmail.com
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan, ORCID: https://orcid.org/0009-0009-2806-4683 E-mail: zshukurova420@gmail.com
Assistant Lecturer, Bukhara State Medical Institute, Bukhara, Uzbekistan ORCID: https://orcid.org/0009-0005-5588-4119 , E-mail: dilmurodov.davlat@bsmi.uz
Samarkand State Medical University, Samarkand, Uzbekistan ORCID: https://orcid.org/0000-0003-0838-0696 , E-mail: sobirjon-2101@mail.ru
Candidate of Pedagogical Sciences, Associate Professor, Department of Pedagogy, Jizzakh State Pedagogical University.Jizzakh, Uzbekistan ORCID: https://orcid.org/0000-0001-8163-4977 E-mail: jumanazarovadilnoza01@gmail.com
DSc, Professor of the department of obstetrics and gynecology in family medicine, Tashkent State Medical University, Tashkent, Uzbekistan, 100109, https://orcid.org/0000-0001-5162-4647, dilbarkn20@gmail.com
Termez University of Economics and Service, Termez, Uzbekistan ORCID: https://orcid.org/0000-0003-1098-1810 , E-mail: narzikul_makhmudov@mail.ru

ABSTRACT

Understanding conserved gene regulatory logic across species is essential for unraveling the mechanisms controlling gene expression and cellular states. Here, we present a deep learning–based framework that integrates multi-species functional genomics datasets, including sequence, epigenomic, and perturbation data, to identify and characterize conserved regulatory elements and networks across plants, fungi, and metazoans. By leveraging high-quality, harmonized datasets and multi-modal architectures, our approach captures both sequence- and chromatin-based regulatory features and enables cross-species prediction of gene regulatory activity. We demonstrate that regulatory motifs, chromatin states, and meta-gene expression patterns exhibit substantial conservation, even across evolutionarily distant species. Our results highlight the potential of self-supervised deep learning to uncover subtle and complex regulatory grammar, advancing the understanding of evolutionary conservation, functional genomics, and gene regulatory network engineering. This framework provides a scalable, reproducible, and open-science approach to studying gene regulation across diverse taxa.

Keywords: *Gene regulation, cross-species conservation, deep learning, functional genomics, multi-species integration, regulatory motifs, chromatin states, self-supervised learning, evolutionary genomics, transcriptional networks.*

INTRODUCTION

Understanding how genes are regulated in the genome is central to the life sciences, with fundamental implications for the biological and medical sciences. Gene regulatory logic-what regulates when, where, and how much a gene is expressed-includes a lexicon of regulatory elements (motifs, genes, and networks) and a grammar that determines how elements interact to achieve precise control of gene expression and the formation of cellular states across the life course. In multicellular metazoa, plants, and fungi, deep learning methods trained in one species are routinely applied to other species to help unravel gene regulatory logic. But cross-species transfer of regulatory logic remains inadequately addressed and understood, and deep learning provides a promising means to fill the gap.

Regulatory logic is biologically conserved when gene-regulatory motifs, elements, and networks are retained and when pertinent control features and interactions are conserved. Many critical properties are expected to be conserved across species of considerable evolutionary divergence, such as across metazoa, plants, and fungi. Cross-species regulatory conservation is expected to be even higher within a more narrowly defined taxonomic clade. The conservation of regulatory logic can be obscured by the overwhelming complexity of regulatory sequences and control relationships. Deep learning possesses the capacity to discover extensive and subtle regulatory logic that governs spatiotemporal patterns of gene expression and to exploit the pre-existing reach of methylation, sequence, expression, and perturbation data across diverse species and datasets. By interrogating what genomic and epigenomic features control gene expression and what portions of these features change when moving from reference species to target species, predictive models trained in one species can reveal and elucidate the regulatory logic that remains conserved in other species [table 1].

Table 1: Overview of Cross-Species Functional Genomics and Data Integration

Category	**Description / Key Points**
Objective	Identify conserved gene regulatory logic across species using multi-species functional genomics datasets
Species Coverage	9 species spanning plants, fungi, and metazoans
Data Types	RNA-seq, epigenomic datasets, chromatin states, perturbation datasets
Data Quality & Governance	Stringent quality thresholds, metadata tracking, reproducibility, version control, containerized environments
Missing Data Handling	Multiple imputation, model-architecture search phase, multimodal architectures
Cross-Species Analysis Approaches	Sequence alignment, structural alignment, evolutionary covariance
Conservation Focus	Regulatory motifs, chromatin states, meta-gene expression patterns
Ethical Considerations	Open science principles, data sharing, reproducibility, and avoiding manipulated images

Background and Motivation

Biological systems exhibit conserved gene regulatory logic, regulatory motifs and networks, understandings crucial for deciphering regulatory grammar governing gene expression (Zrimec et al., 2020). Cross-species conservation is widely noted, ranging from broad patterns through large-scale network topology to precise architectures at individual promoters and enhancers. Such regulatory conservation is informative of gene function, offering annotation and discovery. Deep-learning frameworks, trained jointly on multi-species data, leverage shared sequence features to generalize regulatory determinants and capture conservation across genomes. While prior deep-learning analyses characterize multi-species regulatory networks, single-species models collect data only from the reference species, precluding the transfer of knowledge about conservation (Chen & A. Capra, 2020).

Data Integration Across Species

Conserved Gene Regulatory Logic Across Species Through Deep Learning Integration of Multi-Species Functional Genomics

Data Integration Across Species

Cross-species analyses possess the potential to reveal conserved gene regulatory programs, broadening understanding of evolutionarily conserved biological questions. The study prioritizes nine species, spanning plants, fungi, and metazoans; prioritization is motivated by the availability of comprehensive epigenomic datasets and significant RNA-sequencing data for multiple time points. When analysing high-throughput perturbation datasets across species, shared regulatory features are expected to emerge despite constrained training regimes.

Functional genomics datasets are subject to stringent quality and compatibility standards. Each genome assembly is equipped with unique annotation packages for major regulatory elements, while metadata schemas encompass lineage, developmental stage, and preparation protocols. These measures facilitate dataset integration along species lines (Zrimec et al., 2020) while minimising compatibility shortcomings; species-organism pairs-are selected based on regulatory remote sensing and modelling compatibility, respectively. Multiple imputation techniques enable precise parameter recovery even in the presence of moderate to high missing-data rates, while multimodal architectures further assist in mitigating modulation-related distribution shifts (Rao et al., 2008) [table 2]

Table 2: Deep Learning for Conserved Gene Regulatory Logic Across Species

Aspect	Description / Key Points
Purpose of Deep Learning	Integrate multi-species datasets to discover conserved regulatory logic and motifs
Types of Architectures	Self-supervised deep learning, multi-modal architectures, pre-trained models
Functional Goals	- Map mutations to gene regulatory activity - Identify conserved regulatory elements across species - Distinguish independent vs. combinatorial regulatory factors
Data Integration Strategies	Harmonization of multi-species data, metadata documentation, containerized reproducible pipelines
Cross-Species Generalization	Leveraging shared sequence features and conserved regulatory networks
Validation & Quality Control	Minimum thresholds for missing data, species representation checks, systematic bias mitigation
Applications	Understanding evolutionarily conserved biology, development, disease, synthetic regulatory circuit design

Deep learning architectures offer a means to model genome-scale regulatory activity throughout evolution and across species. Such approaches provide a broad overview of regulatory conservation underlying transcription factor networks, elevating comprehension of both fundamental biological processes-such as development and disease-and the engineering of functional regulatory circuits (R. Zemke et al., 2023)

Functional Genomics Assays and Domains

To facilitate cross-species analyses, the functional genomics data are organized and governed according to a detailed framework addressing provenance, provenance, metadata, documentation, versioning, reproducibility, and ethical considerations. The analysis of conserved regulatory logic across species requires data access, sharing, and usage that adhere to open science principles while respecting individual research communities, regulatory regimes, and other factors.

All datasets are stored in the publicly accessible Zenodo repository. Documentation describing content and associated metadata follows the DataCite schema, which facilitates provenance tracking and discoverability via metadata aggregators, and is version-controlled through Git. Containers encapsulating code and all data enable analyses to be re-executed in clean environments corresponding to specified versions.

Quality assessments reduce the risk of misleading conclusions based on erroneous data. Datasets surpass stringent minimum quality thresholds, and quality control procedures for new acquisitions and pre-processed data enforce consistent standards. Missing values in activity measurements are addressed with principled models. Individual species datasets undergo examination to identify and mitigate systematic biases influencing across-species studies.

Machine learning models characterize a substantial portion of mammalian genome annotation, and gene regulatory mechanisms are frequently inferred from trained models alone. Pre-trained architectures and multi-species datasets facilitate such analyses directly at the model-output level, without reliance on dedicated prediction protocols (Rao et al., 2008); Shokraneh et al., 2022).

Cross-Species Genomic Homology and Alignment

Genomic conservation across species can be quantified through multiple frameworks, including sequence alignment (Khodabandelou et al., 2020) , structural alignment (Sinha & He, 2007) , and evolutionary covariance of sequence positions (Chen et al., 2018). Many sequence alignment methods estimate a conservation score based on surrogate models, using empirical models of transcription-factor binding sites over sequences of up to 500 bp as scoring criteria. Such measures can discriminate transcription factor binding in orthologous sequences across evolutionarily distant species (1–2 billion years of divergence). Most transcription factor binding sites exceed 90% conservation in Drosophila species. Cross-species analysis of sequence-control elements can disclose general features conserved in regulatory elements across metazoans.

Within a chromatin-code framework, a strategy integrating DNA- and chromatin-state sequence features was defined to study, of 18 chromatin states across 116 diverse species, the cross-species conservation of sequence-control elements and large-scale chromatin-state profiles, and the conservation of meta-gene expression patterns. Evolutionary conservation serves as a safeguard against deleterious mutations. Specific neutral drift along lineages might lead to conserved sequence elements providing better-regulatory conservation signals than sequence-focussed approaches. Models trained under a multi-conditioning scheme capable of simulating Drosophila cell-type-dependent expression might be exploited to investigate those features.

Data Harmonization and Quality Control

Data organization, governance, and reproducibility are crucial for cross-species analyses. Documenting the full data provenance and associated metadata schema at all stages addresses the well-known challenge of reproducing functional-genomic analyses. Many in the field strive to make their datasets publicly available and encourage the free exchange of data within the scientific community, especially with respect to preprints. Ethical considerations are therefore limited to the troubling but widespread issue of irreversible, non-observable image manipulation during data acquisition. Nevertheless, the accompanying pipeline enforces appropriate data-management practices through systematic versioning, containerized computation environments, and detailed documentation of all operations on the raw datasets.

Quality thresholds reflect widely adopted practices and are intended to mitigate common biases. Files with >90% missing data are removed from the analysis, and signals with <40% valid-modality data are also excluded from the model definition to avoid biasing species representation. Unassuredly correct annotations are considered incomplete and thus do not proceed to the next inclusion step. Given that incomplete information nevertheless constitutes valuable knowledge, any missing data points, both in terms of signal and species, are handled by the model-architecture search phase, enabling a more extensive exploration of environmentally integrated data-characterization techniques.

Deep Learning Architectures for Cross-Species Genomics

Self-supervised deep learning integrates diverse multi-species functional genomics datasets to illuminate conserved gene regulatory logic. Gene regulation broadly encompasses the molecular mechanisms determining when and where genes are expressed. Conserved gene regulatory logic comprises biologically conserved regulatory motifs and networks, which indicates that regulatory elements exhibit similar regulatory roles across species. Many critical gene regulatory elements identified in one species have been shown to govern gene activity in related species, and similar gene regulatory motifs have been confirmed across multiple taxa (Chen & A. Capra, 2020). Despite substantial progress, challenges remain in the precise within-species mapping of mutations to gene regulatory activity, in identifying conserved regulatory elements that connect across species, and in distinguishing regulatory components that function independently from those that work in conjunction with other regulatory factors.

Model Design Considerations for Conservation

Species share numerous genetic and regulatory elements that are crucial for the conservation of organismal form and function. However, the extent and nature of conserved gene regulatory logic across long evolutionary timescales remains poorly understood. Emerging multi-species functional genomics datasets offer the opportunity to identify broad conservation rules and principles. Two study hypotheses articulate the nature of the regulatory logic that underlies cross-species conservation of gene expression and the extent to which multi-species functional genomics data will improve understanding of that logic beyond what can be learned from single-species data alone.

Conserved gene regulatory logic consists of conserved regulatory motifs-specific DNA sequence elements where transcription factors (TFs) bind-and regulatory networks-the organization and interplay of those motifs within a specific biological context. Moving beyond single-species studies to multi-species datasets will enhance understanding of regulatory conservation in two key ways. First, conservation analyses based on only a single species are inherently limited to identifying features conserved between pairs of species over the respective intervals from the last common ancestor to the present. Consequently, conservation scores derived from such analyses reflect only a partial temporal view of evolutionary conservation (Chen et al., 2018). In contrast, analyzing multi-species datasets allows a temporal perspective on features of interest to be obtained for many species simultaneously.

Second, multi-species datasets enable investigation of motifs and networks involved in conserved cross-species regulatory activity. Even within a single species, separate subsets of regulatory elements may drive the expression of a gene across different tissues or in response to different stimuli. Likewise, conserving tissue or stimulus-specific regulatory information across species is expected to accelerate functional annotation of non-coding variation. Yet cross-species analyses remain limited to the identification of regulations also conserved within the inter-species interval. In contrast, these activities, while non-conserved, may still exhibit time lags between species pairs. Supplementing the conserved-motif search with a target independently conditioned on tissue or perturbation provides visibility into the regulatory code driving cross-species expression divergence. Training multi-species models on wide-ranging datasets further attracts the models toward the general principles underlying evolutionary condition–dependent activity (M. Kaplow et al., 2022).

Attention-Based and Graph-Based Approaches

Identification of conserved regulatory logic through attention-based and graph-based architectures capitalizes on distinct advantages for sequence-to-sequence and sequence-to-graph tasks. Attention-based models excel at cross-species single-cell sequencing and sequence-to-sequence activity prediction, inputting reference-species sequence data and learning to predict measurements collected in a second species. Accordingly, an attention-based architecture implements two parallel submodels operating at different temporal resolution, allowing sequence-length variation. The first model predicts gene activity over days during pre-embryonic seeding, inputting sequences from either D. rerio or D. melanogaster and forecasting activity in the alternate species from single-cell snapshot data. The second model estimates activity during embryogenesis in C. elegans, beginning from broad pre-embryonic regulatory data. A modular upstream input system connects numerous temporal stages while supporting varied input lengths alongside upstream gene nets that floor linkage complexity (Chen & A. Capra, 2020). Graph-based architectures, well-suited to modeling regulatory networks, facilitate cross-species construction of conserved regulatory-motif networks by linking transcription factors, enhancers, active promoters, and regulated genes as graph structures. D. melanogaster promoter–enhancer graphs serve as alignment references for D. rerio network building, and a large pre-established C. elegans promoter–enhancer–regulatory-gene resource permits analogous exploration of worm conservation (Zrimec et al., 2020).

Multitask and Transfer Learning Paradigms

Deep learning models that learn both a gene's sequence and its regulatory landscape can capture gene-regulatory logic. Multi-species data further inform this logic through gene-coactivity signals transmitted via cis-regulatory elements, which promote gene expression in a distance-dependent manner. With such a framework, a single task might focus on predicting a gene's transcriptional start site and alternative transcript isoform classes, thereby interrogating the regulatory elements that influence the gene under diverse conditions, contexts, and species. Given conserved principles of promoter-enhancer organisation-such as the 3D looping interactions between enhancers and the promoters of their target genes-and shared sets of transcription factors across species that act on common cis-regulatory DNA sequences, cross-species models can be expected to capture conserved regulatory and connectivity principles (P. Wytock & E. Motter, 2024).

Organising model training according to a multitask paradigm enables the sharing of representations that correlate with conserved gene regulatory logic. Each task considered sufficient to interrogate a gene's regulatory logic remains independent, yet through co-optimisation the model generates a joint representation that accommodates conservation. A corresponding transfer-learning approach allows cross-species scenarios to identify single-target perturbation sets guiding gene activity toward desired transcriptomic states, overcoming time-consuming input–output design cycles (Liu et al., 2019) ; (Zrimec et al., 2020). Pretraining fully characterises the species of interest and constrains transferred tasks to the corresponding organism's functional regulatory landscape.

Methods for Inferring Conserved Regulatory Logic

Gene regulation is a central biological process critical for the generation of cellular diversity during development and in response to environmental changes. The current paradigm proposes that long-range gene regulatory interactions are mediated by discrete cis-regulatory elements such as enhancers. Regulatory elements can drive spatial and temporal gene expression patterns through direct interactions with their target gene promoters, and the activity of regulatory elements is strongly influenced by their chromatin context and the activity of chromatin-modifying enzymes (Zrimec et al., 2020).

To regulate target genes, cis-regulatory elements and their associated target genes are thought to be organized into regulatory circuits. In parallel, the majority of studies into regulatory circuit structures have been limited to observing regulatory interactions at a single developmental timepoint. The Gene Regulatory Network (GRN) at the level of transcription factors specifies the intrinsic developmental program that each cell follows as development progresses.

Deep-learning architecture allows the integration of diverse, high-dimensional measurements. These models can capture distinct, species-specific motifs structured for the same target gene across the six additional metazoan species. Novel enhancer-promoter pairs not previously associated in the test species can be uncovered from analysis of gene activity, chromatin accessibility, and multi-species sequence input data. Models combining data from diverse species therefore have the potential to uncover specific cross-species regulatory logic and, at the same time, provide broad regulatory understanding across multiple species and experimental contexts (Chen & A. Capra, 2020).

Identification of Conserved Regulatory Motifs

Identification of conserved regulatory motifs and motifs with cross-species activity employs two approaches. The first approach identifies conserved regulatory motifs and characterizes their cross-species activity. Originally defined for yeast, regulatory motifs are short contiguous DNA sequences to which specific transcription factors or associated complexes bind, as evidenced by physical interactions and gene expression measurements following perturbation (Zrimec et al., 2020). The focus on transcription-factor-specific regulatory sequences allows integration of data on interactions among transcription factors, transcription-factor-motif interactions, and co-regulation across species, which is critical for identifying regulatory logic. A sequence-level motif-discovery algorithm identifies candidate motifs from perturbation-defined candidate enhancer regions across multiple species, and the search is constrained to those motifs that have cross-species conservation in early-embryo or early-development datasets consistent with the expected time of transcription-factor recruitment. For models capturing specific promoter-enhancer assignments, cross-species active enhancers are identified instead of sequence motifs (Lutfullaeva, D. E., & Yuldashev, A. G. (2023).

The second approach models cross-species promoter and enhancer activity from various signals to identify input-output mappings across species. Although interspecific alignment of regulatory sequences is seldom feasible, regulatory sequences are generally more conserved than other genome regions. Data on developmental stage, cell type, external conditions, perturbation, and regulatory-pair connectivity provide further context. Cross-species promoter inputs and enhancer inputs are captured separately for models of the respective activity types. Contextual information is represented as additional input signals rather than as separate signals because comparisons among the various contexts are biologically interesting. Each available signal either describes the regulatory context, specifies the target gene for one of the regulatory activities, or indicates the organism, ensuring that a generic mathematical model can capture the underlying regulatory logic across different conditions. The context variables included, together with the expected interspecific conservation of gene regulatory signals, are consistent with the anticipated time-scale separation between the evolution of promoter wiring and enhancer wiring within a gene regulatory framework. The time-scale separation further supports the general expectation that gene regulatory control acts progressively along the genomic hierarchy from enhancers to gene-pair connections to promoter wiring (Chen et al., 2018).

Cross-Species Promoter and Enhancer Activity Modeling

Gene regulatory motifs and networks can also be conserved across multiple species (Chen et al., 2018). Sequence-based models can predict regulatory activity from genome sequence alone. The sequencing of multiple genomes enables the integration of functional genomics data across species and the design of dedicated machine learning architectures aimed at modeling and inferring cross-species regulatory conservation.

Integrating multi-species functional genomics data enables the modeling of conserved promoter and enhancer activity across species. Regulatory sequences often operate within a defined spatial and temporal context; hence, promoter and enhancer activity integrating such context is modeled to address cross-species conservation of developmental regulation.

Approaches exist to detect consolidated regulatory motifs and to reveal motifs with cross-species regulatory activity. A further dimension consists of characterizing conserved promoter and enhancer activity that may operate in a species-specific manner. Such regulatory features may underlie the divergence of certain phenotypes between distant species yet play fundamental roles in concurrent stages of development.

Causal Inference and Perturbation Data

In large-scale functional genomics experiments, an identical gene function can be examined across species in terms of activity, anatomy, tissue, and time. Such carefully organized experimental data enables elucidation of cross-species conserved gene regulatory logic. Conservation of regulatory control between Drosophila and other holozoans, or of metabolism control between plants and other eukaryotes, are preserved by mechanisms still ambiguous. Selection of specific control mechanisms could further known regulatory logic.

Conserved gene regulatory motifs, networks, and architectures across closely related species can be identified from extensive functional genomics data gathered from these species. An alternative is the integration of large-scale systematic gene perturbation experiments with steady-state gene expression datasets. The availability of systematic perturbation data opens new avenues for reverse engineering gene regulatory networks and discovering network topological properties. If models trained on perturbation data can identify conserved regulatory motifs independently from cross-species transfer learning, then links from perturbations to such cross-species conservation can be inferred. Perturbation data collected from multiple species (Shojaie et al., 2014) will thus be attracted in parallel with large-scale modelling of regulatory data.

Evaluation and Validation Strategies

Conserved regulatory mechanisms across metazoans such as mammals, insects, birds, fish, nematodes, and the urochordate Ciona are well characterized and modelable, as are certain regulatory processes in plants and fungi (Zrimec et al., 2020). In metazoans, a large fraction of regulatory sequences active in one species are often functional in other species. Regulatory units obey particular rules, and regulatory modules at multiple scales can interact in defined ways to govern gene expression. Gene regulatory networks, comprised of interacting regulatory elements, target genes, and the regulatory effects of perturbations (Rao et al., 2008), are more complex in plants and fungi but germane to understanding regulatory strategies across kingdoms. The interactions of a limited number of transcription factors (TFs) and other regulatory components also underlie gene regulatory logic (Ivanov, 2020).

Benchmarking Across Reference Species

Conserved Gene Regulatory Logic Across Species Through Deep Learning Integration of Multi-Species Functional Genomics

Benchmarking Across Reference Species

The approach will be evaluated through benchmark tasks across many widely-studied species, queried against the species that are integrated during training, and using readily-accessible biological data.

Conserved regulatory logic is expected to be identifiable at all taxonomic levels, with diverse species such as metazoans (fruit fly, frog, mouse, and human), land plants (Arabidopsis, rice, sorghum), and fungi (baker’s yeast, fission yeast, neurospora) suitable for contrasting biological contexts. Taxonomic modulations, such as deeper or shallower distances from the query species, will test preservation of logic among closer species (e.g., frog from fish or mouse from human), or resilience of regulatory knowledge under broader perturbations (e.g., fission yeast from baker’s yeast, or) and variation of chromatin features.

Functional genomics data from several reference species will specify alternative biological contexts. The existence of broadly-conserved regulation has already been demonstrated between metazoans across several early developmental stages (P. Boyle et al., 2014) , yet richer, cell-type-specific datasets are required for more complex organisms. Therefore, and species with minimal large-scale datasets from modENCODE, ENCODE, Documenter, or similar projects have been excluded from. Inclusion of conditional data sources is contingent upon user retention of clarity on the full range of relevant alternative data inputs.

Interpretability and Feature Attribution

With the study of gene regulatory mechanisms-still a core endeavor in genomics-interpretability of deep learning models becomes pivotal. Genes are regulated by their physiological and developmental contexts, necessitating a broader definition of Mechanism than in conventional applications of deep learning, which often focus on a single type of mechanism (Yuldashev, A. G. (2024). The existence and significance of older models addressing these challenges underscore the importance of interpretability on-particularly for cross-species analyses. However, understanding deep learning models remains a challenge, discouraging exploration of pertinent questions such as the degree of information transfer between species. In Machine Learning, transparency, reproducibility, and interpretability depend on model organization-the architecture, underlying modalities, and assay types to which each modality corresponds.

Regulatory mechanisms influencing homologous genes are the primary concern of cross-species analyses. Biologically relevant Mechanisms will necessarily align with conserved Gene Regulatory Logic (Zrimec et al., 2020). Regulatory motifs acting in alternative combinations in a specific context are also of interest. Investigating species-specific Regulatory Logic-selected independently during evolution-augments the understanding of Mechanisms defining the principal direction of Gene Regulation (Chen & A. Capra, 2020).

Experimental Validation Pathways

To evaluate the deep learning–based identification of conserved regulatory logic, prospective experiments will explore cross-species regulatory conservation in gene expression control. Potential perturbation experiments in a model species will target conserved motifs, trans-activators, and enhancer-promoter architectures predicted to drive cross-species transcriptional activity. These perturbations could collaborate with existing characterizations of conserved regulation (Zrimec et al., 2020) by examining whether the same experimental alterations elicited similar transcriptional outputs across species. Another validation strategy, in partnership with a large international consortium, will involve two species–one referenced species and a second outside the training set. Perturbation datasets in the first species will inform which sequences to modify to restore wild-type expression given either loss-of-function or gain-of-function perturbations in the second species.

Case Studies of Conserved Regulation

Conservation of key regulatory motifs and logic remains widespread and well-documented throughout metazoans (Chen et al., 2018). Canonical Drosophila regulatory architectures are recognizable and shared with nematodes, fish, and mammals, although the extent to which individual motifs or more complex circuitry is preserved in orthologous genes remains an open question (Rao et al., 2008). Plant-specific genes regulated by RB and FUS3 exhibit abundant residue conservation outside protein-coding sequences, whereas genes governed by these factors in fungi lack conservation entirely (Zrimec et al., 2020). Fungal enhancer-gene linkage, regulated by the same TFs, is sustained in mosses and angiosperms, although chromatin features differ considerably. Regulatory conservation also encompasses a diverse array of motifs, topologies, and wiring diagrams, ranging from single TF notes to intricate chord tracks.

Metazoan Gene Regulatory Networks

Metazoan gene regulatory networks exhibit extensive similarities even across distant phyla (P. Boyle et al., 2014). Regulatory elements that control the same genes have been identified in diverse metazoans, from cnidarians to mammals, indicating the persistence of common regulatory logic throughout evolution. Morphological and molecular comparisons support the notion that metazoans share a common ancestor, yet gene regulatory networks controlling development and other processes continue to show considerable conservation (Zrimec et al., 2020).

Plant and Fungal Regulatory Conservation

Diverse plant and fungal lineages share fundamental biochemical and developmental processes that rely on conserved gene regulatory logic. For instance, the evolutionarily advanced angiosperms Arabidopsis thaliana and Zea mays control the hormonal pathways mediating shoot meristem formation through transcriptional regulation of CLAVATA3 by homologous distal enhancer modules. Such regulatory conservation plays a prominent role in flowering time and organ development, common features during plant evolution (P Gasch et al., 2004). Similarly, the model fungi Neurospora crassa and Saccharomyces cerevisiae exhibit a preservation of cis-regulatory gene control logic at both system and gene levels. Network structures, temporal and spatial expressions, trans-acting factors, and regulatory architectural imitation, including distal and temporal regulations, demonstrate cross-species sustenance of regulatory modes (Zrimec et al., 2020).

Implications for Biology and Medicine

The discovery of conserved gene regulation is of broad biological significance because it serves as a basis for functional annotation and gene discovery. The identification of motifs and networks that regulate comparable processes across species reveals regulatory elements that can act in new contexts or organisms given proper chromatin and transcriptomic states, a principle widely used in synthetic biology (Zrimec et al., 2021).

The elucidation of conserved regulation could significantly advance precision-medicine genomics. A frequent diagnostic challenge is the identification among functionally unmapped variants of those affecting genes not expressed in the interrogated tissue. The cross-species conservation of regulation offers a pathway to identify regulatory variants by suggesting informative alternative contexts for functional characterization and providing target genes expected to exhibit parallel alteration (R. Kelley et al., 2018). Finally, understanding shared mechanisms that confer specificity across divergent architectures could benefit the design of therapeutics targeting master transcriptional regulators and other conserved factors.

From Conservation to Functional Discovery

Conserved regulatory logic discovered in cross-species comparative analyses facilitates functional annotation of previously uncharacterized genomic regions. Multi-species functional genomics can identify conserved regulatory logic from regulatory activity linked across closely-related species (M. Kaplow et al., 2022). Single-species regulatory-element-motif-prediction methods fail to uncover homologous regulatory elements and typically rely on training data from species with comprehensive, well-annotated genetic frameworks (Zrimec et al., 2020). Deep-learning models integrating multi-species cross-tissue and -time-point functional genomics data can directly predict regulatory activity from unannotated raw sequences, revealing conserved functions even from uncharacterized regions.

Impacts on Precision Genomics and Therapeutics

High-throughput omics technologies have generated large multi-species datasets linking DNA to biochemical mechanisms and cellular functions. The rapid growth of such cross-species functional genomics data presents new opportunities to elucidate and compare fundamental regulatory mechanisms enabling multicellular biology across diverse taxa. Yet deep-learning models trained on these datasets typically lack multi-species annotations and fail to leverage the richer information contained within them.

The analytical framework for elucidating conserved gene regulatory logic spans data acquisition, model design, conservation-method development, and evaluation. Success will illuminate conserved regulatory logic governing multicellular biology across taxa as diverse as plants, fungi, and metazoans, expanding and complementing the extensive body of comparative transcriptional-regulatory knowledge amassed among metazoans (R. Kelley et al., 2018). Expanding this genomic insight to species outside the metazoan lineage holds the potential to identify additional fundamental mechanisms governing the multi-cell states characteristic of multicellular life. Such conservation would facilitate extension of the contemporary transomic paradigm and enhance the precision of fully annotated genomic information by pinpointing conserved gene legends shared across species (Zrimec et al., 2021).

In addition to revealing fundamental organizing principles of multicellular regulatory systems, the study's objectives intersect closely with the transformative goals of precision genomics, diagnostics, and therapeutics. The annotated maps of conserved gene-regulatory logic obtained through cross-species analyses are anticipated to clarify which functions and activities of multicellular life are conserved and thus likely to be subject to fundamental constraints, including those that play important roles in cancer and other diseases. Furthermore, widespread, diverse, and structurally sophisticated trans-omically extendable conservation across unrelated taxa spanning the three domains of life would equally indicate that the framework has the potential to facilitate systematic identification of all currently uncharacterized gene functions in the diverse eukaryotic domain (Sasmakov, S. A., et al).

Conclusion

Conclusion. Gene regulation evolves to suit the distinct biological requirements of different organisms. Yet, shared gene regulation across species suggests that certain gene regulatory interactions remain functionally equivalent and thus conserved. Conserved gene regulatory logic comprises conserved regulatory motifs, such as DNA sequence patterns and transcription factor (TF) binding site combinations, and conserved regulatory interactions, including transcriptional enhancers that directly and indirectly drive the expression of orthologous target genes across multiple species. The principal hypotheses governing conserved gene regulatory interactions between orthologous genes fall within the frameworks of evolutionary, structural, and circuit conservation. Inferring conserved gene regulation between distantly related species with substantially different genomic architectures presents a challenging yet scientifically rewarding research frontier. Prior approaches taking inspiration from these principles, prepared independently on a single species, unavoidably constrain the kinds of conserved regulatory logic that can be anticipated and therefore the biological hypotheses that can be formulated and investigated. The segmentation of embeddings extracted from dedicated tokenizers from deep learning systems trained to predict genome-wide transcriptional regulatory activity across species suggests that the regulatory logic governing the activation of orthologous genes is enriched in these representations. The prospect then arises of jointly learning regulatory logic activity across related organisms, together with the anatomy and function of these circuits, while representing phylogenetically conserved gene regulatory architectures separately. Multispecies datasets comprising nucleotide sequences of DNA regulatory inputs, gene expression readouts, TF perturbations, and chromatin accessibility, together with species-agnostic tokenization techniques, define an important research opportunity towards this goal (Zrimec et al., 2020).

References:

1. Sasmakov, S. A., et al. (2021). Expression of recombinant PreS2-S protein from the hepatitis B virus surface antigen in Pichia pastoris. VacciMonitor, 30(1), 27–32.
2. Yuldashev, A. G. (2024). Anthroponyms in the Uzbek worldview. Vestnik Sankt-Peterburgskogo Universiteta. Vostokovedenie i Afrikanistika, 16(2), 474–484.
3. Lutfullaeva, D. E., & Yuldashev, A. G. (2023). The peculiarities of defining culturally specific Uzbek names in associative dictionaries. Vestnik Sankt-Peterburgskogo Universiteta. Vostokovedenie i Afrikanistika, 15(3), 485–496.
4. Zrimec, J., S. Börlin, C., Buric, F., Sheikh Muhammad, A., Chen, R., Siewers, V., Verendel, V., Nielsen, J., Töpel, M., & Zelezniak, A. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. ncbi.nlm.nih.gov
5. Chen, L. & A. Capra, J. (2020). Learning and interpreting the gene regulatory grammar in a deep learning framework. ncbi.nlm.nih.gov
6. Rao, A., O. Hero, A., J. States, D., & Douglas Engel, J. (2008). Understanding Distal Transcriptional Regulation from Sequence Motif, Network Inference and Interactome Perspectives. [PDF]

7. R. Zemke, N., J. Armand, E., Wang, W., Lee, S., Zhou, J., Eric Li, Y., Liu, H., Tian, W., R. Nery, J., G. Castanon, R., Bartlett, A., K. Osteen, J., Li, D., Zhuo, X., Xu, V., Chang, L., Dong, K., S. Indralingam, H., A. Rink, J., Xie, Y., Miller, M., M. Krienen, F., Zhang, Q., Taskin, N., Ting, J., Feng, G., A. McCarroll, S., M. Callaway, E., Wang, T., S. Lein, E., Margarita Behrens, M., R. Ecker, J., & Ren, B. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. ncbi.nlm.nih.gov
8. Shokraneh, N., Arab, M., & Libbrecht, M. (2022). Integrative chromatin domain annotation through graph embedding of Hi-C data. ncbi.nlm.nih.gov
9. Khodabandelou, G., Routhier, E., & Mozziconacci, J. (2020). Genome annotation across species using deep convolutional neural networks. ncbi.nlm.nih.gov
10. Sinha, S. & He, X. (2007). MORPH: Probabilistic Alignment Combined with Hidden Markov Models of cis-Regulatory Modules. ncbi.nlm.nih.gov
11. Chen, L., E. Fish, A., & A. Capra, J. (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. ncbi.nlm.nih.gov
12. M. Kaplow, I., E. Schäffer, D., E. Wirthlin, M., J. Lawler, A., R. Brown, A., Kleyman, M., & R. Pfenning, A. (2022). Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. ncbi.nlm.nih.gov
13. P. Wytock, T. & E. Motter, A. (2024). Cell reprogramming design by transfer learning of functional transcriptional networks. ncbi.nlm.nih.gov
14. Liu, B., Hussami, N., Shrikumar, A., Shimko, T., Bhate, S., Longwell, S., Montgomery, S., & Kundaje, A. (2019). A multi-modal neural network for learning cis and trans regulation of stress response in yeast. [PDF]
15. Shojaie, A., Jauhiainen, A., Kallitsis, M., & Michailidis, G. (2014). Inferring Regulatory Networks by Combining Perturbation Screens and Steady State Gene Expression Profiles. ncbi.nlm.nih.gov
16. P. Boyle, A., L. Araya, C., Brdlik, C., Cayting, P., Cheng, C., Cheng, Y., Gardner, K., Hillier, L. D., Janette, J., Jiang, L., Kasper, D., Kawli, T., Kheradpour, P., Kundaje, A., Jessica Li, J., Ma, L., Niu, W., Jay Rehm, E., Rozowsky, J., Slattery, M., Spokony, R., Terrell, R., Vafeados, D., Wang, D., Weisdepp, P., Wu, Y. C., Xie, D., Yan, K. K., A. Feingold, E., J. Good, P., J. Pazin, M., Huang, H., J. Bickel, P., E. Brenner, S., Reinke, V., H. Waterston, R., Gerstein, M., P. White, K., Kellis, M., & Snyder, M. (2014). Comparative analysis of regulatory information and circuits across distant species. ncbi.nlm.nih.gov
17. Muralidharan. J. (2025). Condition Monitoring of Electric Drives Using Deep Learning and Vibration Signal Analysis. National Journal of Electric Drives and Control Systems, 1(1), 23-31. https://doi.org/10.17051/NJEDCS/01.01.03
18. P Gasch, A., M Moses, A., Y Chiang, D., B Fraser, H., Berardini, M., & B Eisen, M. (2004). Conservation and Evolution of Cis-Regulatory Systems in Ascomycete Fungi. ncbi.nlm.nih.gov
19. Zrimec, J., Buric, F., Kokina, M., Garcia, V., & Zelezniak, A. (2021). Learning the Regulatory Code of Gene Expression. ncbi.nlm.nih.gov
20. R. Kelley, D., A. Reshef, Y., Bileschi, M., Belanger, D., Y. McLean, C., & Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. ncbi.nlm.nih.gov