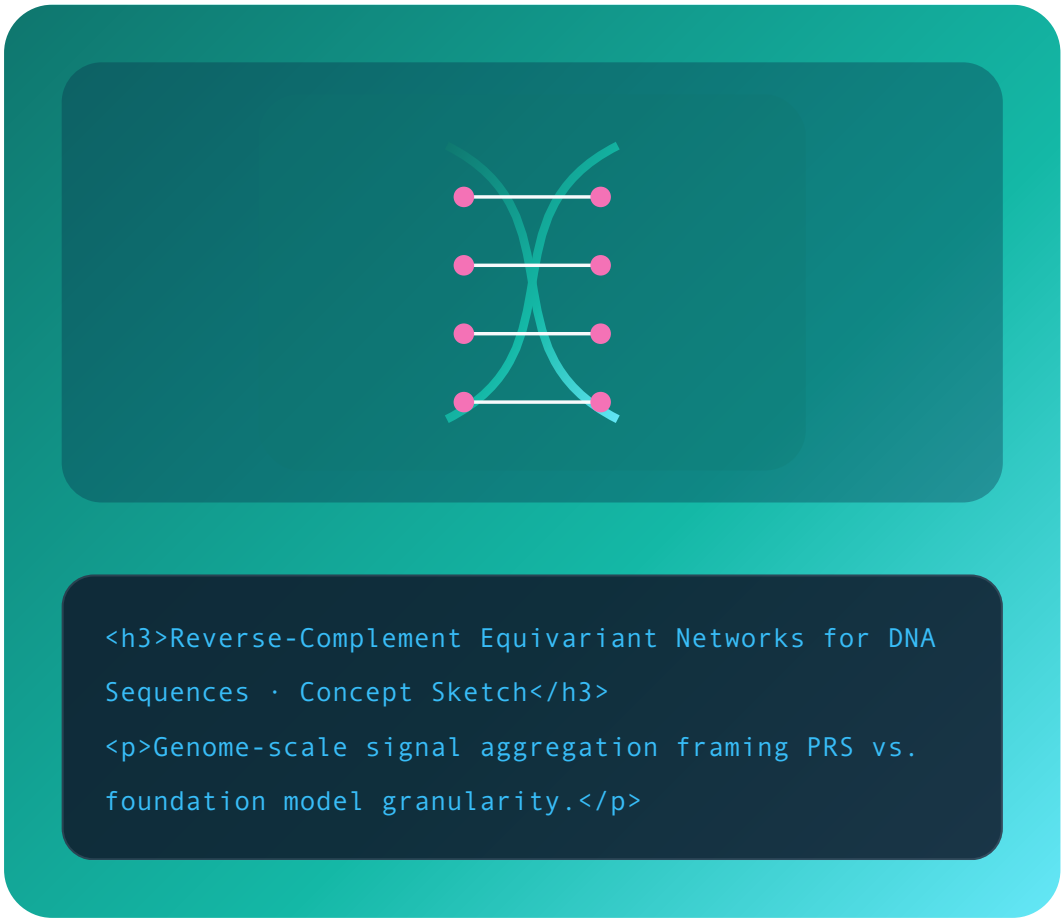

Reverse-Complement Equivariant Networks for DNA Sequences

Authors: Vincent Mallet, Jean-Philippe Vert

Year: 2021

Venue: 35th Conference on Neural Information Processing Systems (NeurIPS)

1. Classification

- **Domain Category:**
  - **Genomics FM.** The paper develops neural network architectures that incorporate reverse-complement (RC) symmetry for DNA sequence analysis, specifically for tasks like transcription factor binding prediction and regulatory element identification.
- **FM Usage Type:**
  - **Core FM development (architectural).** The main contribution is characterizing all possible RC-equivariant layers and proposing new architectures beyond existing RC-parameter sharing (RCPS) methods, rather than proposing a complete foundation model.
- **Key Modalities:**
  - Single-modality DNA sequence (nucleotide-level, with k-mer embeddings as an alternative to one-hot encoding).

---

2. Executive Summary

This paper addresses a fundamental symmetry in DNA sequences: the reverse-complement (RC) property, where a DNA fragment can be equivalently described by two RC sequences depending on which strand is sequenced. While previous work (RCPS) proposed one specific RC-equivariant CNN architecture, it remained unknown whether other architectures exist that could potentially improve performance. Using the theory of equivariant representations and steerable CNNs, the authors characterize the complete set of linear translation- and RC-equivariant layers, showing that new architectures exist beyond RCPS. They introduce irreducible representation (irrep) feature spaces that allow different types of equivariant layers, discuss RC-equivariant pointwise nonlinearities adapted to different architectures, and propose RC-equivariant k-mer embeddings as an alternative to one-hot nucleotide encoding. Experimentally, they show that the new architectures

(particularly irrep-based models with optimized hyperparameters) outperform existing RCPS models on protein binding prediction tasks, with improvements of similar magnitude to the gap between RCPS and non-equivariant models. This paper demonstrates how to systematically incorporate biological symmetries into neural architectures using group theory, and shows that exploring the full space of equivariant architectures can yield practical benefits beyond the first proposed solution.

---

3. Problem Setup and Motivation

- **Scientific / practical problem**
  - DNA sequences have a fundamental symmetry: a double-stranded DNA fragment can be sequenced as two equivalent reverse-complement sequences, and many genomic tasks (e.g., protein binding, regulatory element identification) are RC-invariant or RC-equivariant.
  - The goal is to build neural networks that respect this symmetry by design, rather than relying on data augmentation or post-hoc averaging.
  - Specifically, they want to:
    - Characterize all possible RC-equivariant architectures (not just RCPS).
    - Design new architectures that may outperform existing ones.
    - Understand what nonlinearities and embeddings are compatible with RC-equivariance.
- **Why this is hard**
  - **Limited architectural exploration:**
    - RCPS was the only known RC-equivariant CNN architecture, but it was unclear if alternatives existed.
    - Without a systematic characterization, it’s impossible to know if better architectures are possible.

- **Group theory complexity:**
  - RC-equivariance requires understanding group actions (translation + RC operations form a semi-direct product group Z ⋊ Z2).
  - Different group representations lead to different equivariant layers, and not all are obvious.
- **Nonlinearity constraints:**
  - Not all activation functions are compatible with all equivariant representations.
  - Some representations only allow odd functions (e.g., tanh), while others allow any nonlinearity (e.g., ReLU).
- **K-mer embeddings:**
  - K-mers improve performance but need to be encoded in an RC-equivariant way, which is non-trivial (some k-mers are their own reverse complement).

---

4. Data and Modalities

- **Datasets used**
  - **Binary classification tasks:**
    - Transcription factor binding prediction for three TFs (CTCF, MAX, SPI1) using ChIP-seq data from GM12878 cell line.
    - Sequences labeled as binding vs. non-binding sites.
  - **Sequence prediction tasks:**
    - Base-pair resolution TF binding profile prediction for four TFs (OCT4, SOX2, NANOG, KLF4) in mouse embryonic stem cells using ChIP-nexus data.
    - Outputs are position-wise binding scores.
- **Modalities**
  - Single modality: **DNA sequence** at nucleotide resolution (A/C/G/T).

- Outputs are binary labels (binding/non-binding) or continuous profiles (binding scores per position).

- **Preprocessing / representation**
  - **One-hot encoding:** Each nucleotide (A, C, G, T) encoded as a 4-dimensional one-hot vector.
  - **K-mer embeddings (proposed):** Overlapping k-mers (k=1,2,3,4) encoded in an RC-equivariant way, with special handling for self-complementary k-mers.
  - **Sequence length:** Varies by task (typically hundreds to thousands of base pairs).

---

5. Model / Foundation Model

- **Model Type**
  - **Convolutional Neural Networks (CNNs)** with RC-equivariant layers.
  - Based on steerable CNNs framework for group-equivariant architectures.
- **Is it a new FM or an existing one?**
  - **Architectural innovation, not a complete FM.** The paper characterizes and proposes new RC-equivariant layer types, but does not build a full foundation model. It focuses on architectural components that could be used in DNA foundation models.
- **Key components and innovations**
  - **Characterization of all RC-equivariant linear layers:**
    - Uses group representation theory to show that any RC-equivariant layer corresponds to a representation ρ of Z2.
    - Two main types:
      - **Regular representation:** Used by RCPS, allows any pointwise nonlinearity.

        - **Irreducible representation (irrep):** New type with “+1” and “-1” channels, where “+1” channels are RC-invariant and “-1” channels flip sign under RC. Only odd nonlinearities (e.g., tanh) allowed on “-1” channels.
    - **RC-equivariant pointwise nonlinearities:**
        - Theorem characterizing which nonlinearities are compatible with which representations.
        - Regular representation: any nonlinearity (ReLU, sigmoid, etc.).
        - Irrep with both “+1” and “-1” channels: odd functions (tanh) on “-1” channels, any function on “+1” channels.
    - **RC-equivariant k-mer embeddings:**
        - Encodes k-mers in a way that respects RC symmetry.
        - Handles self-complementary k-mers (e.g., “AT” is its own RC) by using a blend of regular and irrep representations.
    - **RC-equivariant batch normalization:**
        - For irrep spaces: enforces zero mean on “-1” channels, scales variance appropriately.
- **Training setup**
    - **Architecture:** Stack of RC-equivariant convolutional layers followed by pooling and classification/regression heads.
    - **Loss:** Binary cross-entropy for classification, appropriate regression loss for profiles.
    - **Hyperparameters:** Optimized via validation set (k-mer length k, ratio a/(a+b) for irrep channels).

6. Multimodal / Integration Aspects (If Applicable)

- **Not applicable.** This paper focuses on unimodal DNA sequence modeling with architectural innovations for incorporating RC symmetry. It does not integrate multiple modalities.

---

7. Experiments and Results

Main findings

- **New architectures outperform RCPS:**
  - Best equivariant model (irrep-based with optimized hyperparameters) significantly outperforms RCPS on all three binary classification tasks (CTCF, MAX, SPI1).
  - Improvement from RCPS to best equivariant is similar in magnitude to improvement from standard (non-equivariant) to RCPS.
  - On sequence prediction tasks, results are mixed: irrep models outperform RCPS in low-data regime but underperform on full dataset (possibly due to k-mer blurring at nucleotide resolution).
- **Hyperparameter sensitivity:**
  - **K-mer length:** k=3 gives best performance, significantly outperforming k=1,2,4.
  - **Irrep ratio a/(a+b):** Optimal at a/(a+b) = 0.75 (75% “+1” channels, 25% “-1” channels). Performance degrades when a=0 (all “-1” channels) or a/(a+b)=1 (all “+1” channels, equivalent to standard model).

- **Low-data regime:**
  - Equivariant architectures show larger advantages over non-equivariant models when training data is limited (1,000 sequences vs. full dataset).
  - Gap between best equivariant and standard widens from ~0.3% to ~1% AuROC in low-data regime.
- **Ensembling:**
  - Ensemble of two equivariant models outperforms single equivariant models.
  - Post-hoc averaging (averaging predictions on x and RC(x)) helps non-equivariant models but is less effective than architectural equivariance.

Ablation studies

- **Regular vs. Irrep representations:**
  - Regular representation (RCPS-style) with k-mer encoding performs well.
  - Irrep representation with optimized a/(a+b) ratio can outperform regular representation.
  - Choice depends on task and data regime.
- **K-mer encoding:**
  - K-mer embeddings improve performance over one-hot nucleotide encoding in equivariant architectures.
  - k=3 is optimal across tasks.

8. Strengths and Limitations

Strengths

- **Systematic characterization:**
  - First paper to characterize the complete space of RC-equivariant layers, not just propose one architecture.
- **Theoretical rigor:**
  - Uses group representation theory to provide principled understanding of equivariant architectures.
- **Practical improvements:**
  - New architectures outperform existing RCPS models, demonstrating that exploring the full equivariant space is valuable.
- **Comprehensive coverage:**
  - Addresses linear layers, nonlinearities, embeddings, and batch normalization in an equivariant framework.
- **Code availability:**
  - Implementation available in Keras and PyTorch for practical use.

Limitations

- **Not a complete foundation model:**
  - Focuses on architectural components rather than a full pretrained FM.
  - Does not address long-range dependencies or large-scale pretraining.
- **Limited to CNNs:**
  - Does not address RC-equivariance in Transformer-based or other architectures (though principles could extend).
- **Task-specific:**
  - Evaluated only on protein binding prediction; performance on other genomic tasks (e.g., variant effect prediction, gene expression) not shown.

- **Hyperparameter sensitivity:**
  - Optimal k-mer length and irrep ratio need to be tuned per task, which adds complexity.
- **Mixed results on profile tasks:**
  - Irrep models underperform RCPS on full-dataset profile prediction, suggesting task-dependent architecture selection.

Open questions / future directions

- **Extension to Transformers:**
  - How to incorporate RC-equivariance into attention-based models (DNABERT, Nucleotide Transformers)?
- **Long-range dependencies:**
  - Can RC-equivariant architectures scale to very long sequences (100k+ bp) like Caduceus or HyenaDNA?
- **Foundation model integration:**
  - How to combine RC-equivariant layers with large-scale pretraining objectives?
- **Other biological symmetries:**
  - Can similar methods handle other symmetries (e.g., translation, rotation in 3D structures)?
- **Theoretical understanding:**
  - Why does a/(a+b) = 0.75 work best? What is the optimal ratio for different tasks?

---

9. Context and Broader Impact

Relation to other work

- **Compared to RCPS (Shrikumar et al., 2017):**
  - RCPS proposed the first RC-equivariant CNN using regular representation and parameter sharing.

- This paper shows RCPS is just one point in a larger space of equivariant architectures.
- New irrep-based architectures can outperform RCPS.

- **Compared to Caduceus (Schiff et al., 2024):**
  - Caduceus uses Mamba SSMs with explicit RC-equivariance for long-range DNA modeling.
  - This paper focuses on CNNs for shorter sequences; principles could inform long-range equivariant models.
- **Compared to DNABERT-2 / Nucleotide Transformers:**
  - These Transformer-based models don't enforce RC-equivariance architecturally (rely on data augmentation).
  - This paper's methods could potentially be adapted to Transformer architectures.
- **Connection to group-equivariant CNNs:**
  - Builds on steerable CNNs (Cohen & Welling, 2017) and group-equivariant CNN theory.
  - Applies general principles to the specific case of DNA sequences.

Broader scientific and practical impact

- **Systematic approach to biological symmetries:**
  - Demonstrates how group theory can guide architecture design for biological data.
  - Provides a template for incorporating other symmetries (e.g., in protein structures, RNA).
- **Improves DNA sequence modeling:**
  - Better architectures for tasks like TF binding prediction, regulatory element identification.
  - Could be integrated into foundation models for genomics.
- **Theoretical contribution:**
  - Characterizes the complete space of RC-equivariant layers, providing a foundation for future work.

Open questions for future research

- **Transformer integration:**
  - How to make attention mechanisms RC-equivariant?
- **Long-range modeling:**
  - Can RC-equivariant architectures scale to megabase contexts?
- **Other symmetries:**
  - What other biological symmetries should be encoded (e.g., codon translation, RNA secondary structure)?
- **Foundation model design:**
  - How to combine RC-equivariance with large-scale pretraining (e.g., in DNA language models)?

---

10. Key Takeaways

1. **Biological symmetries matter:**
   DNA has inherent symmetries (reverse-complement) that should be encoded in model architectures, not just learned from data.

2. **Group theory guides architecture design:**
   Using group representation theory, you can systematically characterize all possible equivariant architectures, not just guess at solutions.

3. **Exploring the full space pays off:**
   The first proposed equivariant architecture (RCPS) was not optimal; exploring the full space of equivariant layers led to better models.

4. **Different representations enable different nonlinearities:**
   Regular representation allows any activation function, while irrep representation requires odd functions on “-1” channels. This constraint affects model expressivity.

5. **Hyperparameter tuning is crucial:**
   The ratio of “+1” to “-1” channels in irrep representations and k-mer length significantly affect performance and need careful tuning.

6. **Equivariance helps most in low-data regimes:**
   Architectural priors (like RC-equivariance) provide larger benefits when training data is limited.

7. **K-mer embeddings can improve performance:**
   Encoding k-mers (rather than single nucleotides) in an equivariant way can boost accuracy, but requires careful handling of self-complementary k-mers.

8. **Ensembling still helps:**
   Even with equivariant architectures, ensembling multiple models improves performance, though architectural equivariance is more efficient than post-hoc averaging.

9. **Task-dependent architecture selection:**
   Different equivariant architectures (regular vs. irrep) may be optimal for different tasks (classification vs. profile prediction).

10. **This is foundational work:**
    Understanding how to systematically incorporate symmetries into neural architectures is essential for building better models for biological data, and could inform design of DNA foundation models.