C. Origami

Research Question

The paper addresses how chromatin organization determines cell-type-specific gene expression, focusing on overcoming the limitations of experimental methods like Hi-C in terms of cost and scalability for studying chromatin interactions.

Main Benefit

The study presents *C.Origami*, a multimodal deep neural network that integrates DNA sequence and cell-type-specific genomic features (e.g., CTCF binding, chromatin accessibility) to predict chromatin organization de novo. This enables efficient in silico experiments for genetic screening, providing a scalable alternative to experimental techniques.

Alternative Methods

The paper compares C.Origami with other models like Akita, DeepC, and Orca. These rely either on sequence features or epigenomic signals but lack the combined multimodal approach, limiting their accuracy in predicting cell-type-specific chromatin organization.

Architecture

- **Inputs:** DNA sequence, CTCF ChIP-seq, and ATAC-seq signals.
- **Model Design:** A multimodal encoder-decoder architecture with:
  - Two 1D convolutional encoders for DNA and genomic features.
  - A transformer module for long-range interactions.
  - A 2D convolutional decoder to output Hi-C-like contact matrices at an 8,192-bp resolution.
- **Outputs:** Hi-C interaction matrices predicting chromatin topology, such as TADs and loops.

Metrics Used

1. **Insulation Score Correlation**:
   - Measures how well the model predicts insulation scores, which indicate how well a region of the genome is "insulated" or isolated from neighboring regions.
   - High Pearson and Spearman correlations (~0.95) with experimental data indicate strong predictive accuracy.
2. **Distance-Stratified Interaction Correlation**:
   - Evaluates how well the model predicts chromatin interactions at varying genomic distances.
   - This is crucial because chromatin interactions vary in strength and frequency based on distance (e.g., short-range loops vs. long-range interactions).
3. **Observed/Expected Hi-C Map Correlation**:
   - Compares predicted interaction frequencies to experimental ones, normalized for biases in Hi-C data.
   - This metric is helpful for capturing broad trends in chromatin organization.
4. **Mean Squared Error (MSE)**:
   - Assesses the pixel-level difference between predicted and experimental Hi-C matrices.
   - Lower MSE indicates a closer match to the experimental data.
5. **Area Under the Receiver Operating Characteristic Curve (AUROC)**:
   - Used for loop-calling, which identifies specific interactions like promoter-enhancer loops.
   - AUROC scores of 0.92 (for the top 5,000 predicted loops) demonstrate the model's ability to detect biologically relevant features.

Potential Additional Metrics

1. **Normalized Mutual Information (NMI)**:
   - Measures the similarity between predicted and experimental TAD structures. It captures how well the model reconstructs domain-level organization.
2. **Precision-Recall AUC**:
   - While AUROC is robust for balanced datasets, Precision-Recall AUC could better evaluate performance in detecting rare chromatin loops or interactions.
3. **Jaccard Index**:
   - Useful for comparing the overlap of predicted and experimental contact domains, such as TAD boundaries.
4. **Structural Similarity Index (SSIM)**:
   - Common in image analysis, SSIM could quantify how well the overall structure of predicted Hi-C matrices matches experimental ones.
5. **Reproducibility Score**:
   - This would assess the consistency of model predictions across replicates or noisy inputs, which is particularly relevant for experimental genomics data.

Validation

The validation of *C.Origami* involved both /quantitative metrics and qualitative comparisons to assess its accuracy and robustness across different datasets and cell types:

1. **Data Setup**:
   - Hi-C data from human cell lines (e.g., IMR-90, GM12878, K562, H1-hESC) were used.
   - Chromosomes were split into training, validation, and testing sets, ensuring no overlap between sets. For example, chromosome 10 was used for validation, while chromosome 15 was held out for testing.
   - The model also extended to interspecies validation with mouse chromatin data to test generalization.
2. **Validation Goals**:
   - To confirm that the predicted Hi-C contact matrices accurately represent chromatin interactions.
   - To test the model's ability to generalize to unseen cell types or species.
3. **Evaluation Methods**:
   - Predicted Hi-C matrices were compared to experimental ones using correlation metrics.
   - Loop-calling experiments assessed how well the model could detect specific chromatin interactions like enhancer-promoter loops.
   - Genome-wide analyses evaluated structural features like TADs and insulation scores.
4. **Cell-Type Specificity**:
   - De novo predictions were tested on new cell types (e.g., GM12878) to assess cell-type-specific chromatin organization.
   - Predictions were robust across diverse genomic contexts, including differential chromatin structures and conserved regions.

What is GRAM?

Summary: GRAM is a key interpretability tool in C.Origami that highlights the genomic regions most critical for chromatin folding predictions. It is particularly useful for generating hypotheses about regulatory elements, which are further explored using more robust methods like in silico perturbations or experimental validations.

GRAM (*Gradient-weighted Regional Activation Mapping*) is a gradient-based saliency method used in *C.Origami* to identify genomic regions (cis-elements) that significantly influence chromatin organization. It builds on the idea of saliency maps commonly used in computer vision to interpret neural networks by highlighting important input features that drive predictions.

How GRAM Works

1. **Gradient Computation**:
    - GRAM calculates the gradient of the model's prediction with respect to the input feature values (e.g., DNA sequence, CTCF ChIP-seq, ATAC-seq signals).
    - The gradient indicates how sensitive the output (e.g., predicted Hi-C contact values) is to changes in a specific input region.
2. **Regional Attribution**:
    - Instead of focusing on individual nucleotide-level features, GRAM aggregates the gradients over predefined genomic regions (e.g., TAD boundaries or specific chromatin features).
    - This approach assigns an *importance score* to each region, quantifying its contribution to the predicted chromatin organization.
3. **Visualization**:
    - GRAM generates heatmaps or attribution maps that highlight the genomic regions most responsible for specific chromatin interactions, such as enhancer-promoter loops or TAD boundaries.

How GRAM is Used in *C.Origami*

1. **Identifying Critical Cis-Elements**:
    - GRAM is applied to the input features (DNA sequence, CTCF, ATAC-seq) to identify regions where small perturbations lead to significant changes in predicted Hi-C contact matrices.
    - These regions are inferred as critical cis-elements driving chromatin organization.
2. **Biological Insights**:
    - GRAM highlights known features like CTCF-binding sites and accessible chromatin regions as major contributors to chromatin folding.
    - It also identifies previously uncharacterized elements, providing new insights into genome organization.
3. **Validation with Experimental Data**:
    - Regions identified by GRAM as critical for chromatin structure were cross-validated with experimental Hi-C and ChIP-seq data to confirm their biological relevance.
4. **Integration with In Silico Genetic Screening (ISGS)**:
    - GRAM is used as a preliminary tool in the ISGS pipeline to prioritize regions for systematic perturbation (e.g., in silico deletion or modification).
    - This speeds up the identification of impactful regions by narrowing the focus to high-gradient areas.

Advantages and Limitations

Advantages:

- **Interpretability**: Provides a clear visualization of which regions are important for chromatin organization.
- **Efficiency**: Quickly identifies key regions without requiring full in silico perturbations across the entire genome.

Limitations:

- **Window Shifts**: GRAM’s sensitivity can vary with small shifts in the input window, making it less stable for fine-grained interpretations.
- **Gradient Saturation**: For regions with extreme signal values (e.g., very high ATAC-seq peaks), gradients may saturate, reducing interpretability.

Alternative/Complementary Approaches

While GRAM is useful, it is often paired with other methods in *C.Origami*, such as:

1. **Attention Weights from the Transformer Module**:
    - Provide a complementary, robust measure of regional importance by averaging attention scores across layers and heads.
    - More stable than GRAM for noisy or sparse input signals.
2. **Impact Scores in ISGS**:
    - Quantifies the change in predictions after systematic perturbations, providing a direct measure of the functional importance of regions.

ISGS stands for **In Silico Genetic Screening**.

It is a framework used in *C.Origami* to systematically and quantitatively assess how individual DNA elements (cis-regulatory elements) contribute to 3D chromatin organization. By simulating genetic perturbations, such as deletions or modifications of specific regions, ISGS evaluates the functional importance of these elements in shaping chromatin interactions.

1. What is the primary advantage of *C.Origami* over traditional Hi-C experiments?

- **Answer**: *C.Origami* allows high-throughput, cost-effective, and scalable in silico prediction of chromatin organization, reducing reliance on expensive and time-consuming experimental methods like Hi-C.

2. What are the key input features used by *C.Origami* to predict chromatin organization?

- **Answer**: DNA sequence (one-hot encoded), CTCF binding profiles (ChIP-seq), and chromatin accessibility data (ATAC-seq).

3. How does the transformer module in *C.Origami* contribute to the model's performance?

- **Answer**: The transformer module enables long-range information exchange, which is critical for modeling interactions between distant genomic regions.

4. Define the term "impact score" in the context of ISGS.

- **Answer**: The impact score quantifies the effect of perturbing a genomic region on the predicted Hi-C contact matrix, calculated as the mean absolute difference between the matrices before and after the perturbation.

5. How does GRAM identify important cis-regulatory elements?

- **Answer**: GRAM calculates gradients of the model's output with respect to input features and aggregates them over genomic regions to determine their importance in chromatin organization.

6. What metric is used to evaluate how well *C.Origami* predicts insulation scores, and why is this metric important?

- **Answer**: Pearson and Spearman correlation coefficients are used. These metrics are important because insulation scores reflect the separation of chromatin regions, a key feature of 3D genome organization.

7. Name one limitation of GRAM as mentioned in the paper.

- **Answer**: GRAM's results can be unstable when input windows are shifted or due to changes in random seeds.

8. Describe a scenario where *C.Origami* is used to study disease-specific chromatin organization.

- **Answer**: In T-ALL (leukemia), *C.Origami* identified a loss of insulation at the CHD4-insu element, correlating with increased CHD4 expression and new chromatin interactions that promote leukemia cell proliferation.

9. How does *C.Origami* outperform models like Akita and DeepC?

- **Answer**: *C.Origami* integrates multimodal inputs (DNA sequence + genomic features) and uses a transformer module, resulting in higher accuracy in metrics like insulation score correlation and loop detection.

10. Suggest an alternative metric that could be used to evaluate *C.Origami* and explain its relevance.

- **Answer**: The Jaccard Index could be used to measure the overlap between predicted and experimental TAD boundaries, providing insight into how well the model identifies domain-level organization.