

4Dゲノム状態の理解と可視化を支援するデータベース開発

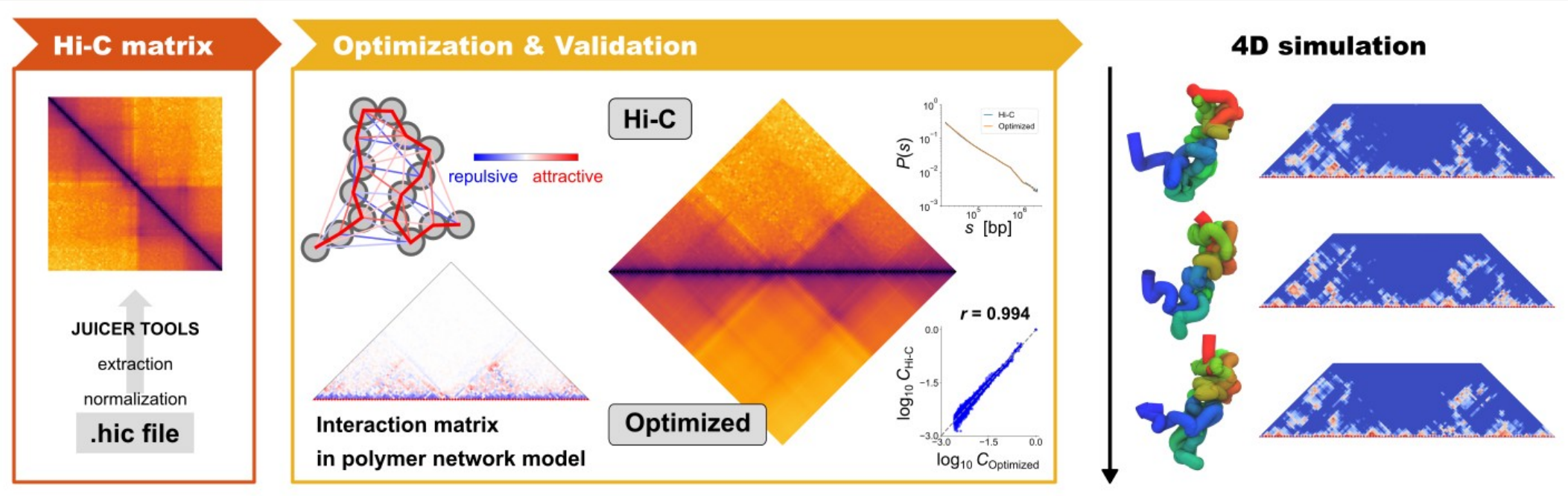
中島 広樹, 萩原 蒼也, 糸賀 裕弥, 大浪 修一, 新海 創也

理化学研究所 生命機能科学研究センター



背景

- Hi-C法により、細胞状態に応じた**3次元（3D）ゲノム構造**の理解が進展
- 膨大なコンタクト配列データを扱う**データエコシステム**は、**米国4D Nucleomeプロジェクト**を中心に整備
- 今後は世界的にHi-Cデータ蓄積フェーズへ
- 一方、**Hi-Cは化学固定を前提**とし、生細胞内のクロマチン動態をどの程度反映しているかは不明
- 生細胞データと整合する物理学的解釈基盤**が必要
- 我々は、Hi-Cコンタクトマップを**99%以上の精度**で再現可能な高分子物理モデル推定手法**PHI-C（ファイシー）**を開発
- PHI-Cにより、2次元Hi-Cデータを**動的3D構造**や**4Dゲノム状態マップ**へ変換可能



Shinkai et al. NAR Genom Bioinform 2020; Shinkai et al. Bioinformatics 2022

目標

- 公共Hi-Cデータを対象に、**PHI-Cによる4Dゲノム状態解析**を網羅的に実施
- 4Dゲノム状態の理解と可視化**を支援するデータベースを構築

現状の課題

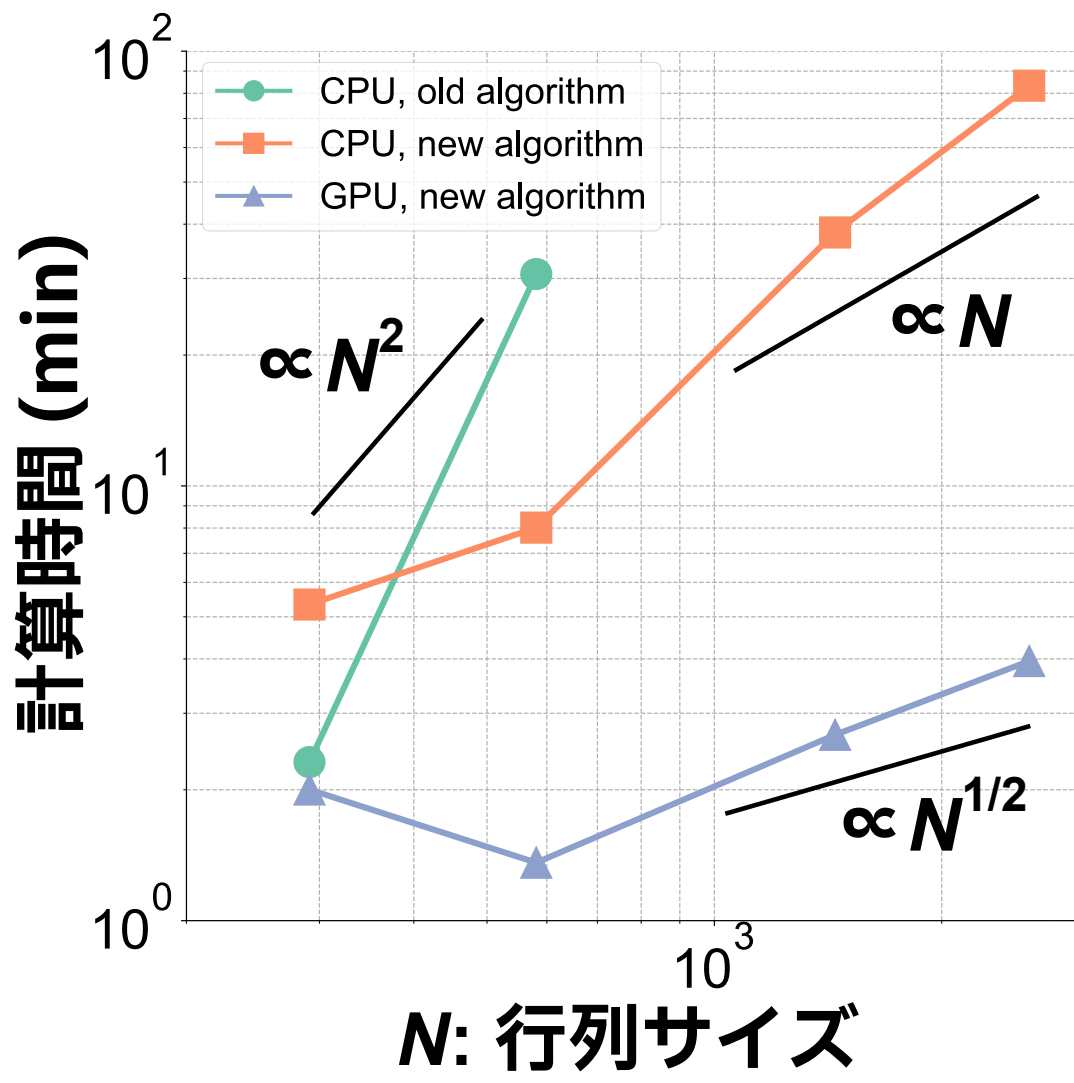
- 最適化アルゴリズム**は行列サイズに対し**N²オーダー**で計算量が増加し、計算効率がボトルネック
- 4Dゲノム状態マップ**や**3D動態シミュレーション**は、従来は別々のソフトウェアでしか可視化できず、**統合環境が不足**

大きな進展となる進捗結果

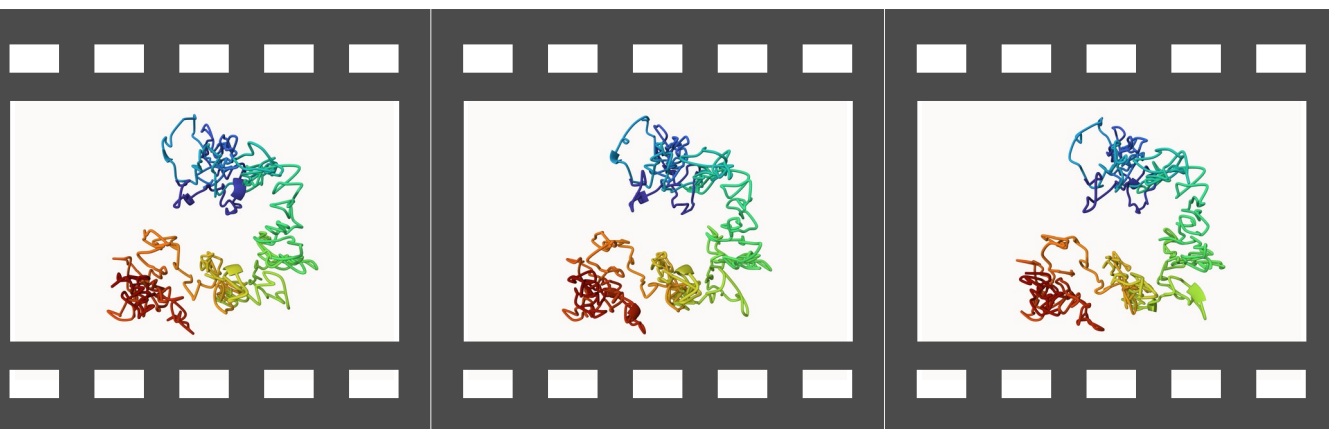
- 最適化アルゴリズムの**理論整備**と**GPU実装**により、高速化と並列化を実現し、実効的に**N^{1/2}オーダー**のスケーリングを達成
- Zarr形式**を導入し、大規模行列データへの効率的アクセスと**複数マップの同時可視化**を可能に
- Mol***を組み込み、**3D動態シミュレーション**を**同一ブラウザで統合表示**

1. PHi-C最適化アルゴリズムの高速化とGPU並列化

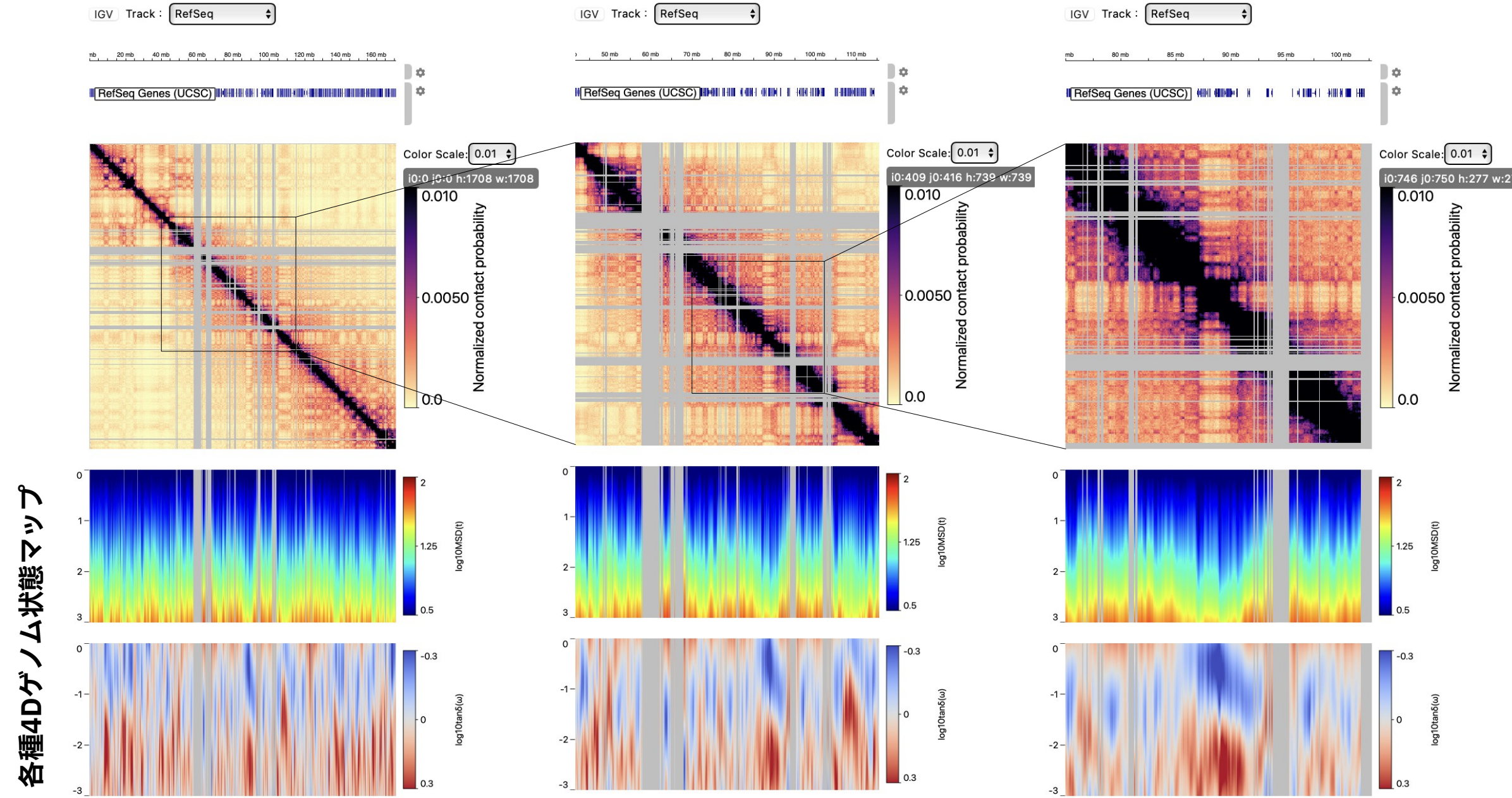
ヒト細胞株GM12878 第8染色体 CPU: Intel(R) Xeon(R) Gold 6132 2.60GHz GPU: Tesla V100-PCIe-32GB



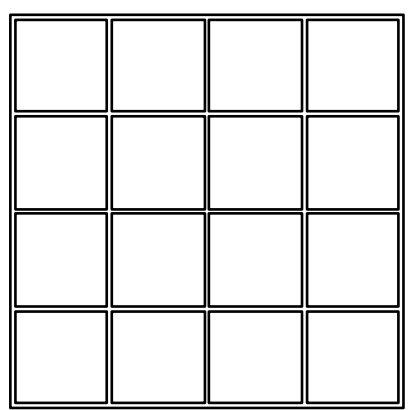
3. Mol*による3D動態シミュレーションの統合可視化環境



2. Zarr形式の導入と複数の4Dゲノム状態マップの同時可視化



各種4Dゲノム状態マップ



- Zarr形式に変換**：大規模配列データを効率的に格納
- チャンク化**：必要部分だけを高速に読み書き
- 5000×5000行列**への高速アクセスを実現

今後の課題と目標

- PHi-C最適化アルゴリズムの精度向上**（目標：99.9%以上の一致度）
- 公共Hi-Cデータベースへの網羅的適用**
- 既存ゲノムデータベースとの連携**を視野に、**包括的かつ直感的なUI**を設計
- 2026年度上半期：β版公開、同年10月：正式版リリース**

References

- David Sehna, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, Alexander S Rose: Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures, *Nucleic Acids Research*, 2021; 10.1093/nar/gkab31.
- Zarr Developers. (2023). Zarr Specification Version 3 (ZEP0001). Zarr Specification Project. Retrieved from <https://zarr-specs.readthedocs.io/en/latest/v3/core/>

This work is licensed under CC BY 4.0.

