DLScaff: Deep Learning and Hi-C data for chimeric contigs detection

Alexis Mergez, Raphaël Mourad, Matthias Zytnicki

▶ **To cite this version:**

Alexis Mergez, Raphaël Mourad, Matthias Zytnicki. DLScaff: Deep Learning and Hi-C data for chimeric contigs detection. JOBIM (JOURNÉES OUVERTES EN BIOLOGIE, INFORMATIQUE ET MATHÉMATIQUES) 2023, Jun 2023, Toulouse, France. hal-04147306

HAL Id: hal-04147306
https://hal.science/hal-04147306

Submitted on 30 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

DLScaff : Deep Learning and Hi-C data for chimeric contigs detection

Alexis MERGEZ, Raphaël MOURAD and Matthias ZYTNICKI

MIAT, Toulouse INRAE, 31326 CEDEX, Castanet-Tolosan, France

Error free contig

→ Aligned with **unique** chromosome

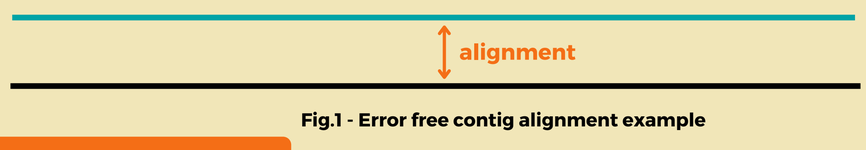


Fig.1 - Error free contig alignment example

Chimeric contig

→ Aligned with **multiple** chromosomes

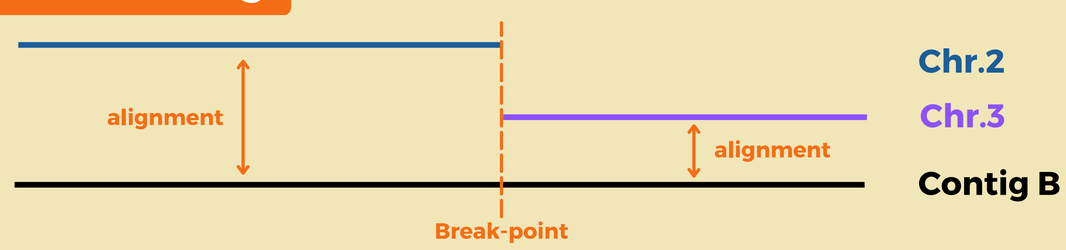


Fig.2 - Chimeric contig alignment example

Hi-C contact matrix

- HiC protocol captures genome to genome interactions
- Close sequences are also spatially close
- **Low** counts regions (see below) indicates **low** mid to long range interactions
- **Low** mid to long range interactions **is a pattern of chimeric contigs**

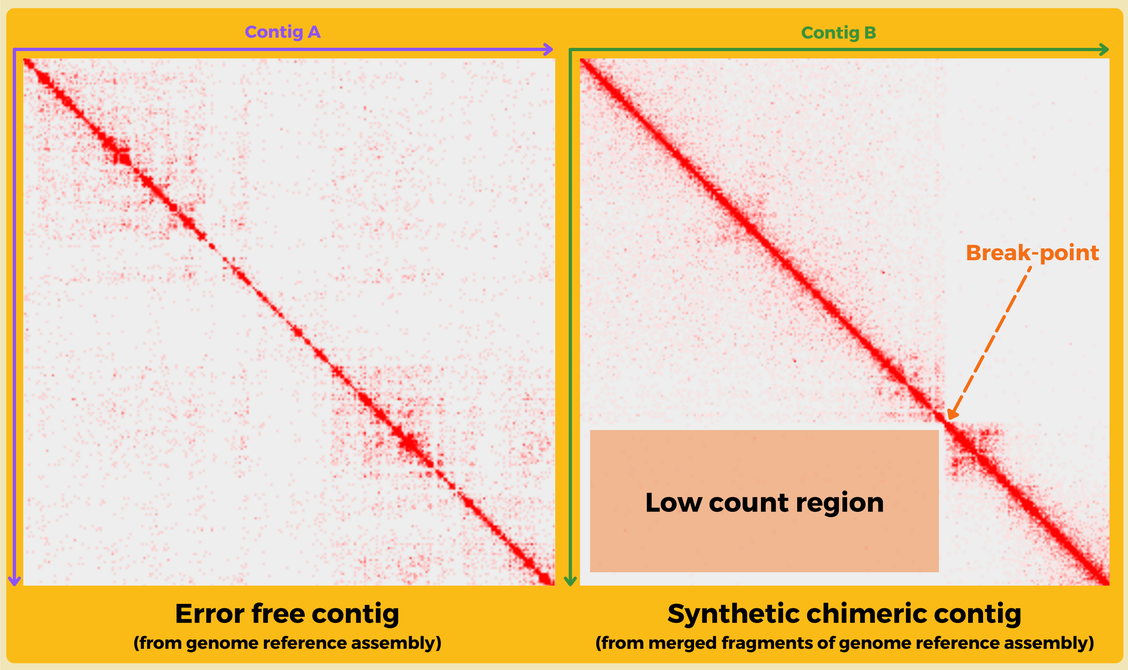
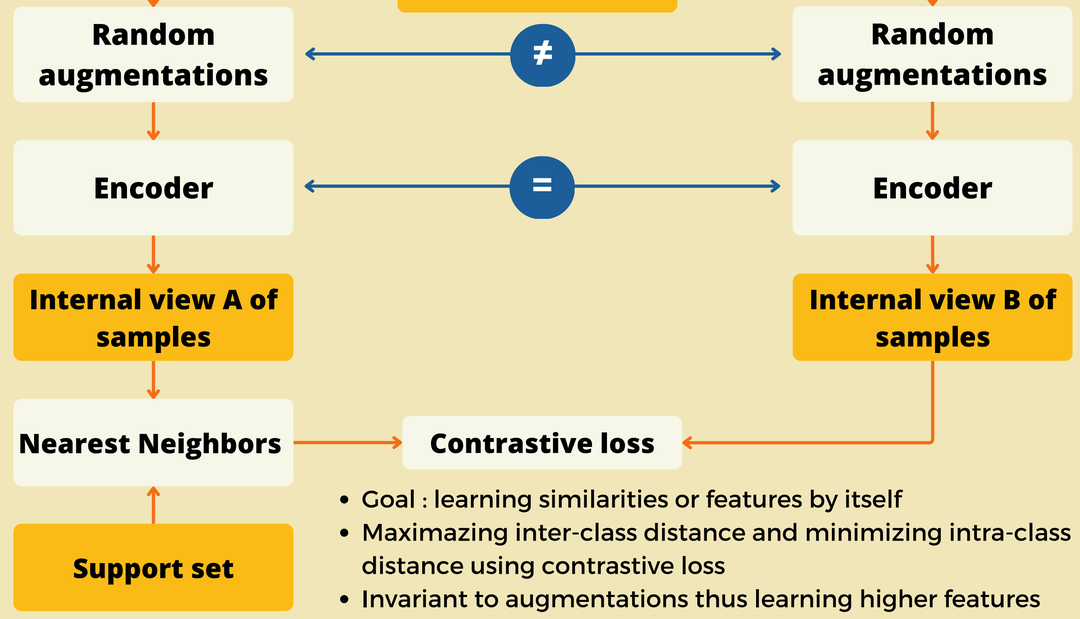


Fig.3 - Comparison between error free and chimeric contig Hi-C contact matrix

Nearest Neighbor Constrastive Learning of visual Representations Deep Learning model



- Goal : learning similarities or features by itself
- Maximazing inter-class distance and minimizing intra-class distance using contrastive loss
- Invariant to augmentations thus learning higher features

Fig.4 - NNCLR whole model architecture overview

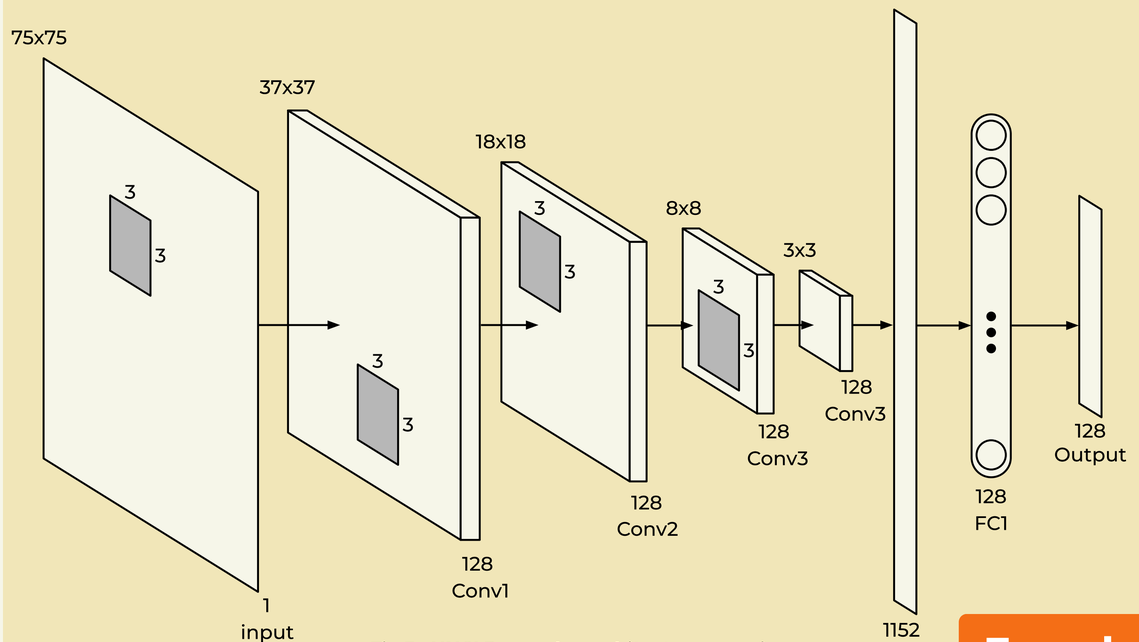


Fig.5 - NNCLR encoder architecture overview

Encoder

Test methodology

- NNCLR and a simple Convolution Neural Network (CNN) tested against YaHS, one of the best available tool
- YaHS search for drops in linkages number
- Three training datasets :
  - ASM : HG002 assembly using hifiasm
  - SYN : synthetic chimeric contigs from HG002
  - SEA : ASM and SYN
  - Negative matrices from reference assembly
- Chimeric contigs are detected using alignments to reference genome
- All models are tested on YaHS synthetic test set with a five-fold cross-validation

Results

- CNN model constantly underperforms or is, at most, on par with others
- NNCLR outperforms YaHS on AUPR, AUROC and GSCORE and is at least on par on accuracy, balanced accuracy and F1-score
- Bad labelisation may be the determinant factor in models bad results seen on ASM
- CNN model seems not to be able to learn higher features compared to NNCLR
- Bigger datasets should reduce variability between folds for NNCLR
- NNCLR model seems to provide better detection than YaHS. Tweaking the model could improve results

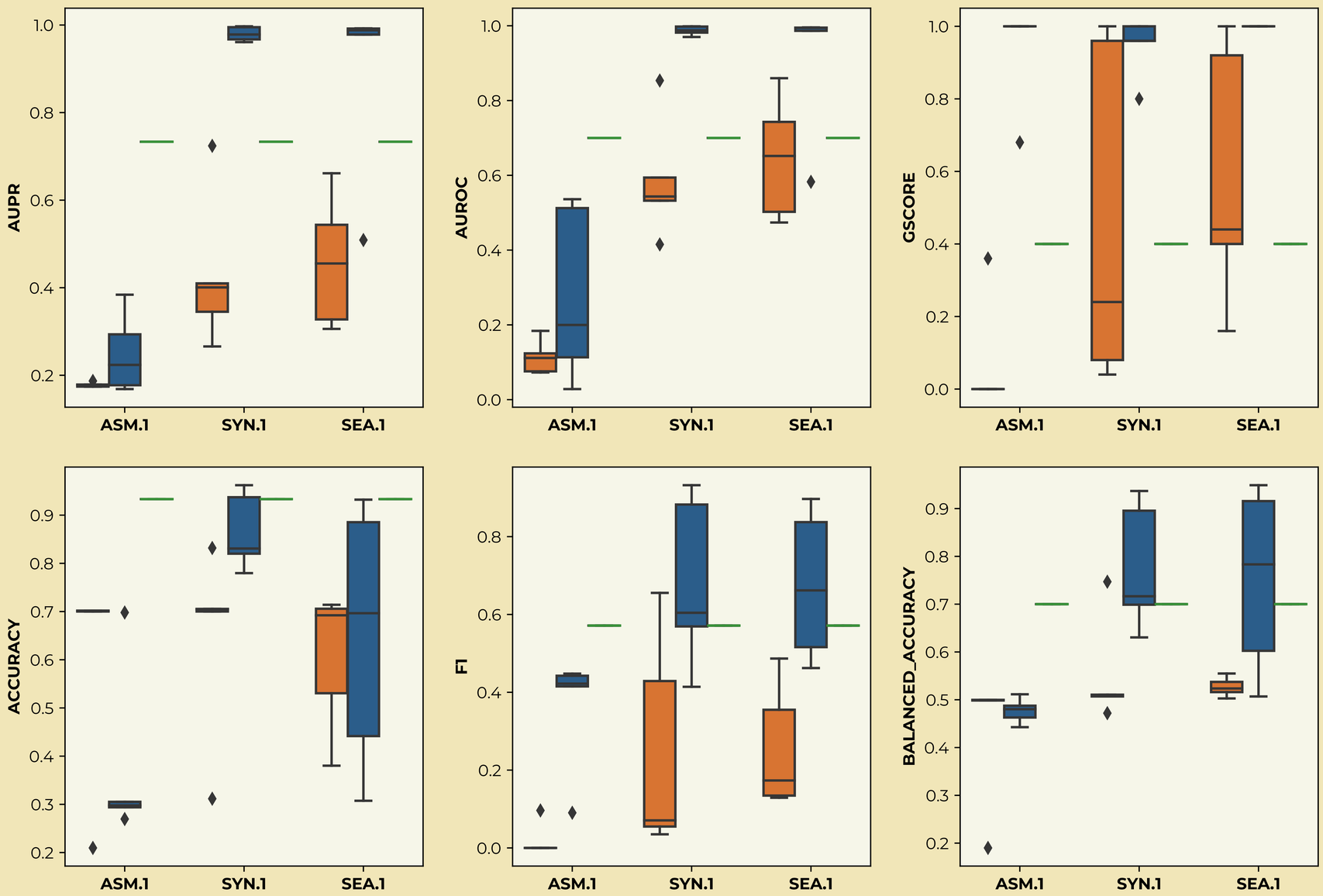


Fig.6 - NNCLR, CNN and YaHS results on YaHS synthetic dataset