Protein structure alignment and genome structure alignment

contact map overlap methods applied to Hi-C chromatin interaction data

Cellular Genome Architecture

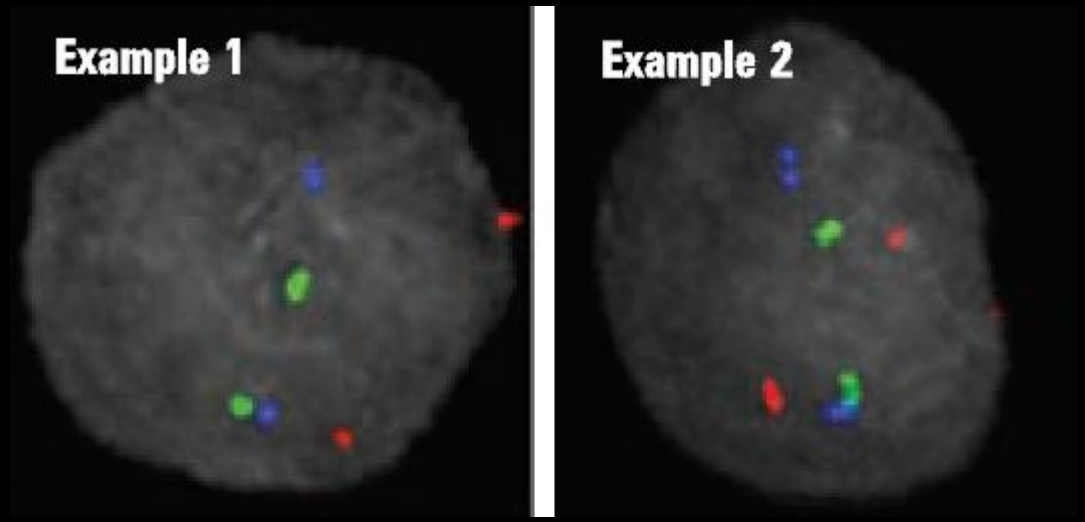
3 billion bases in the human genome
2m stretched end to end
Cell nucleus: 6μm

How do you fit all the DNA?
- Transcription
- Modification
- Repair
- Replication

All processes still need to work!

Microscopy studies

- First way to study chromatin
- Observation of metaphase chromosomes
- Fluorescence in Situ Hybridization (FISH)



- Gene dense together
- Gene poor together
- Looping chromatin for gene regulation

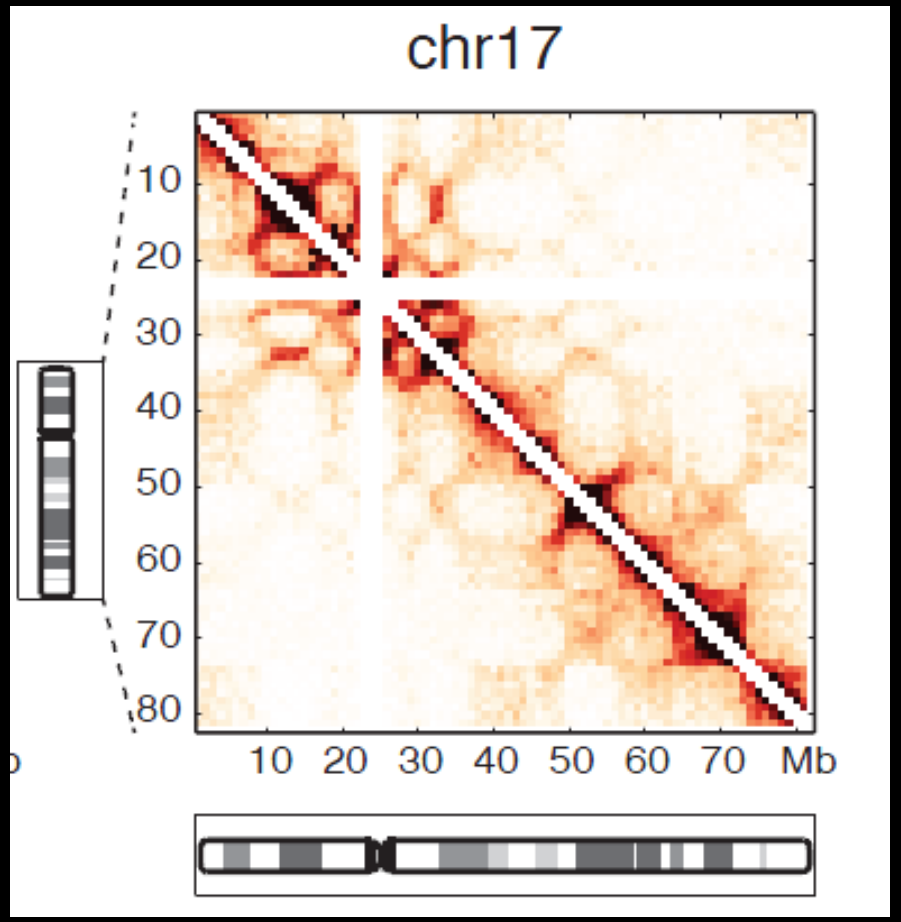
Chromatin Conformation Capture

- Modern microbiology methods
- High throughput sequencing
- Capture what chromatin is close in 3D space

3C	4C	5C	Hi-C	ChiA-PET
"Chromatin Conformation Capture"	"Chromatin Conformation Capture on Chip"	"Chromatin Conformation Capture Carbon Copy"	"Chromatin Conformation Capture High Throughput Sequencing"	"Chromatin Immunoprecipitation + 3C"

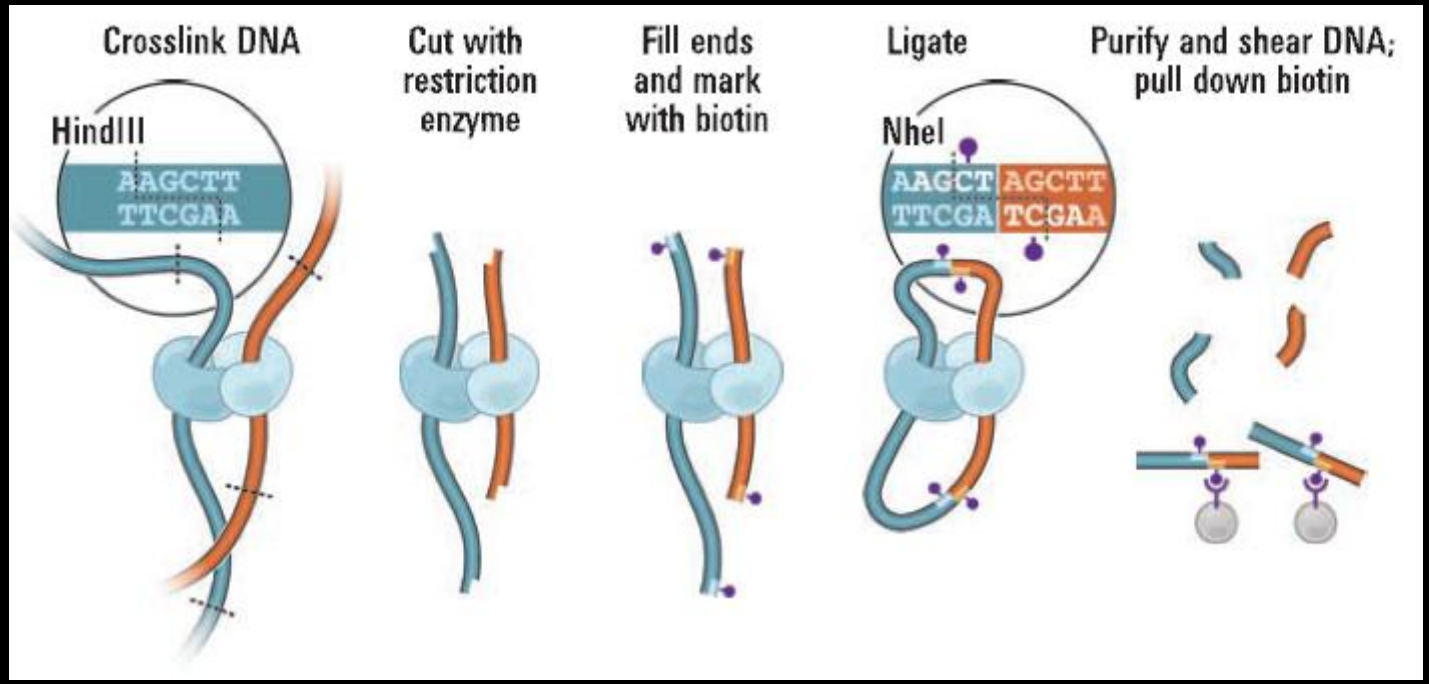
Hi-C

- "All-by-all" assay of chromatin structure
- Conducted on cell populations
- End result: matrix of interaction frequencies
- Interchromosomal
- Intrachromosomal



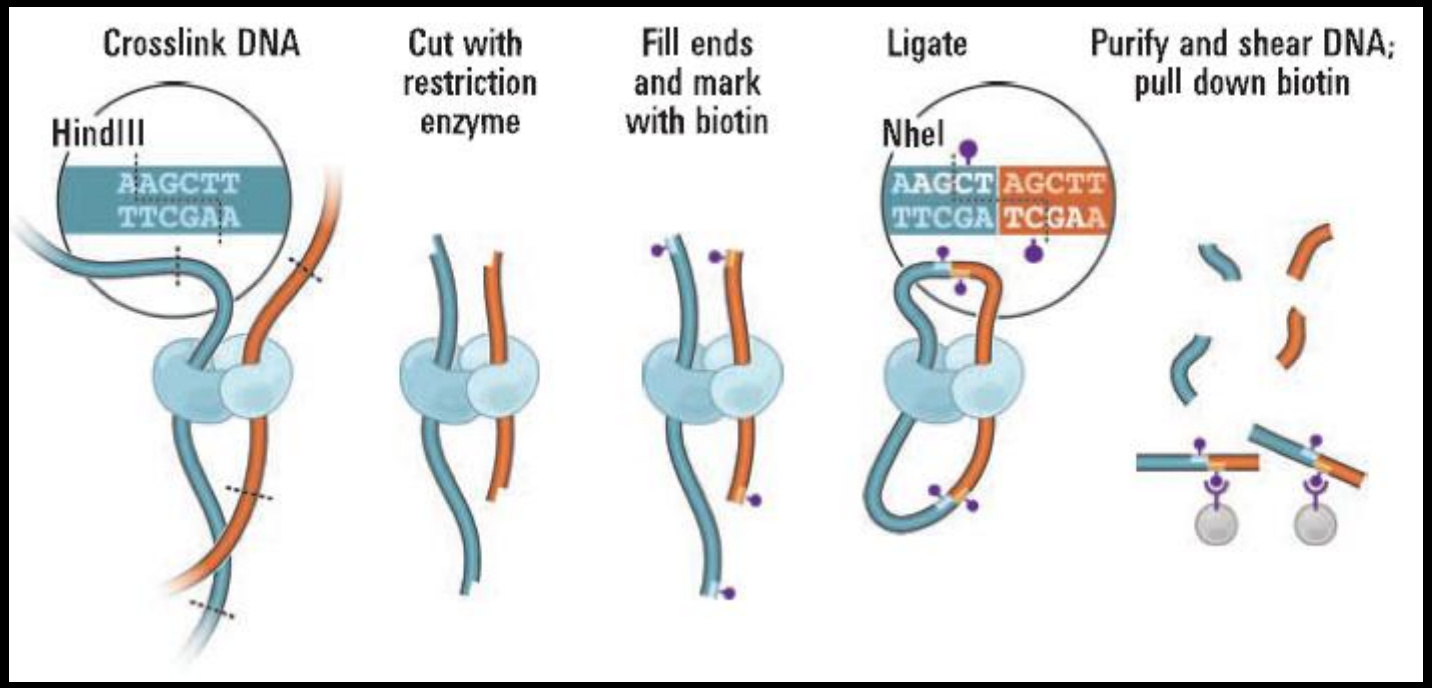
Method

1) Crosslink DNA/protein with formaldehyde
2) Cut with restriction enzyme up/downstream
3) Fill overhang, including biotin



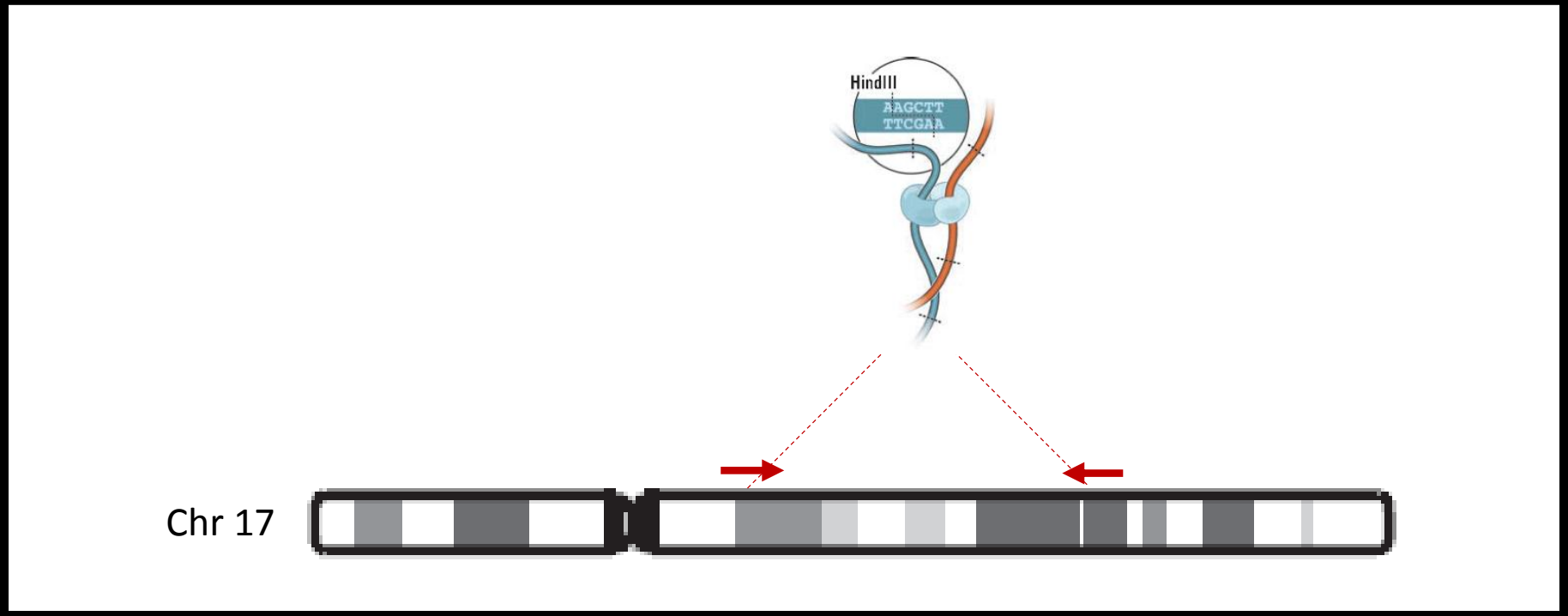
Method (cont)

4) Blunt end ligation to form circles
5) Randomly shear
6) Select for biotin
7) Paired end sequencing



Data Processing

- Align paired end reads to reference genome
- Process at resolution
- Make a heatmap!



Hi-C has been used to study…

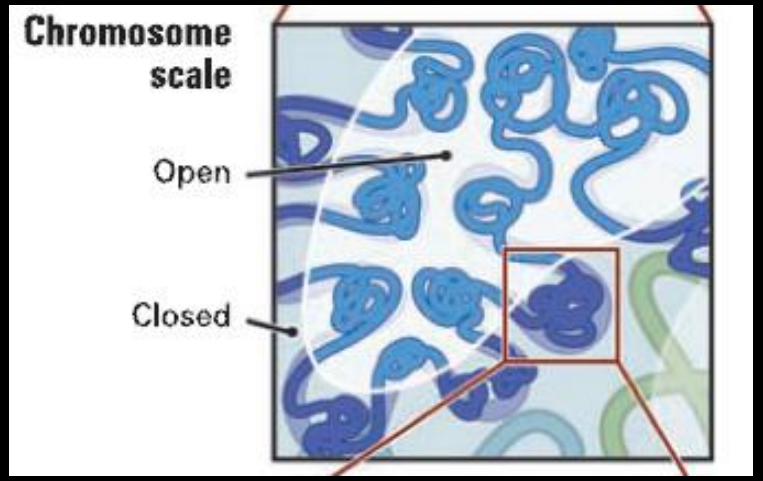
- Regulation of gene expression
- Cell differentiation
- Organization of mitotic chromosomes
- Chromatin structure in Progeria
- Chromatin structure in aging and senescence

My lab (Nicola Neretti)

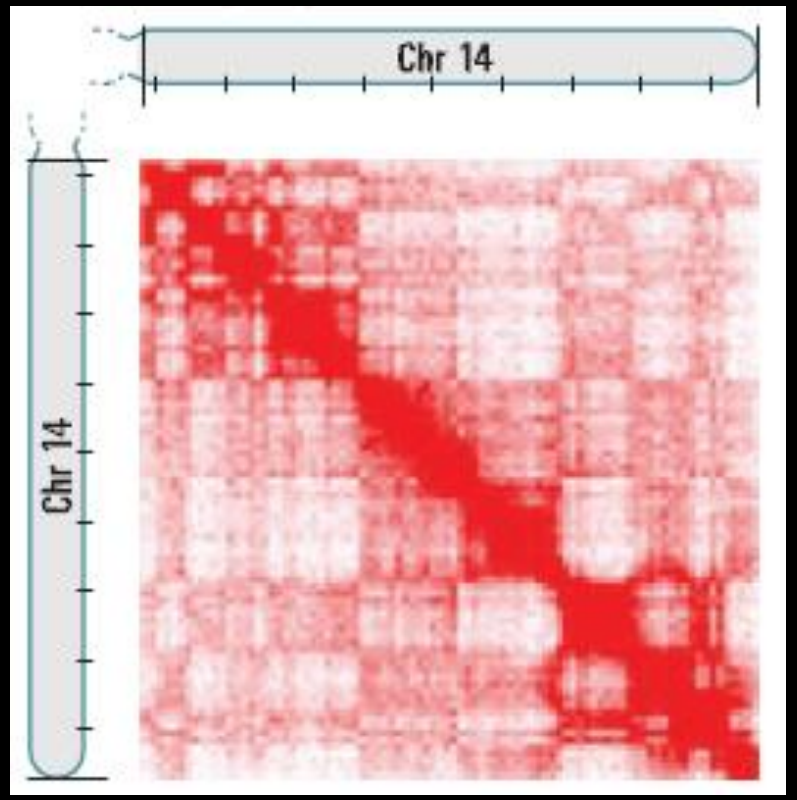
<u>2 universal features of genome architecture…</u>

A/B compartments

- Chromosomes sections occupy distinct domains
- Several megabases in size
- "A" regions interact
- "B" regions interact
- A and B *don't* interact
- Gene rich / gene poor mentioned earlier



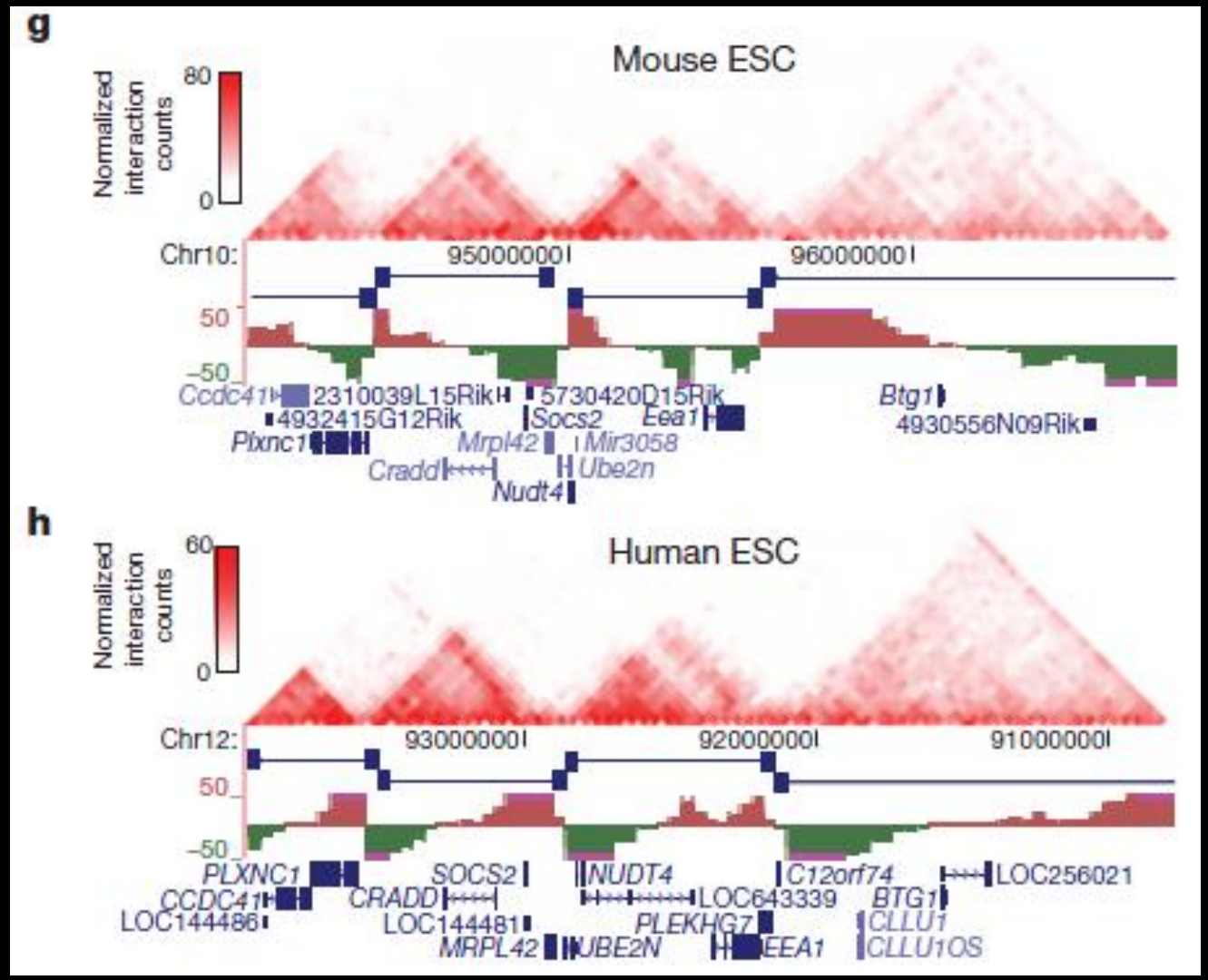
(Lieberman-Aiden et al. 2009)



(Lieberman-Aiden et al. 2009)

Topologically associating domains

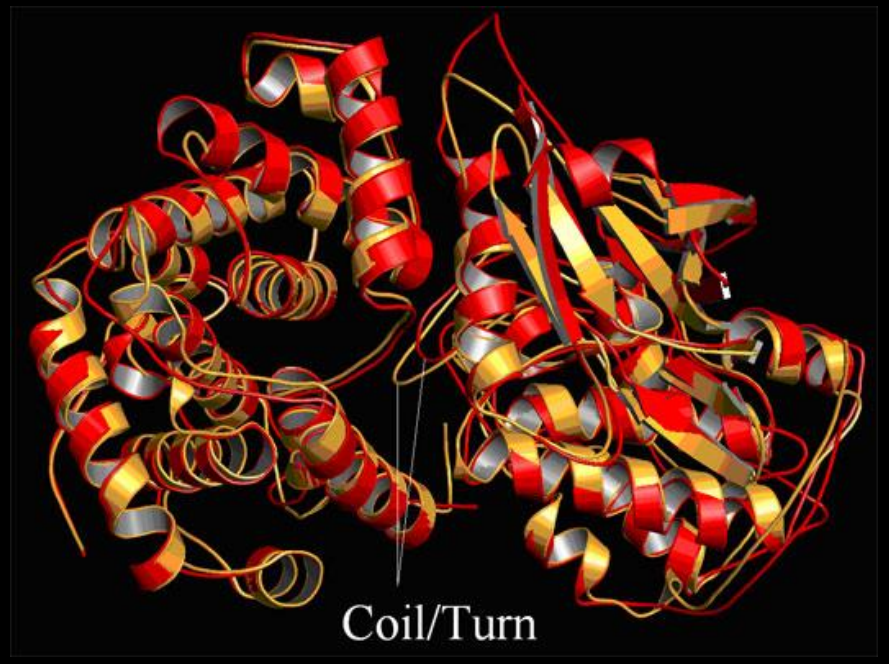
- Sub-megabase resolution
- Repeated structures
- Consistently associated with proteins (CTCF)
- Conserved across cell line and species (!)



(Dixon et al. 2012)

Protein structure alignment

How similar are two protein structures?
Are there conserved domains?
How can we align similar structures?



3D model of a wheat cyclin protein complex (gold) superimposed onto the human CDK2-cyclin-A complex (red)
http://www.jic.ac.uk/staff/graham-moore/wheat_meiosis.htm

Do similar structures have similar function?
How is structure conserved in evolution?

Protein Structure Alignment Methods

- Similarity measure
- Function to optimize

Four are common

- Root Mean Square Deviation (RMSD)
- Distance map similarity
- Universal similarity metric
- Contact Map Overlap

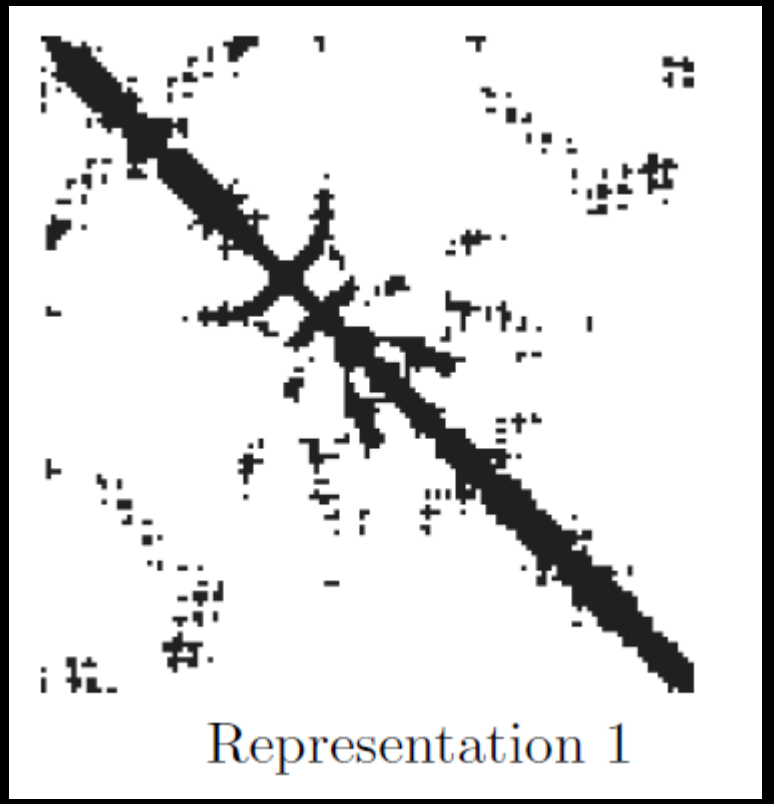
Why Max CMO?

- No linear alignment required
- Can be solved exactly
- Best for highlighting similar domains

What is a contact map?

- Extract a *contact map* from structure
- 2D representation of 3D
- *N x N* binary matrix (graph)
  (i,j) = 1 if i, j are *"in contact"*
  0 otherwise

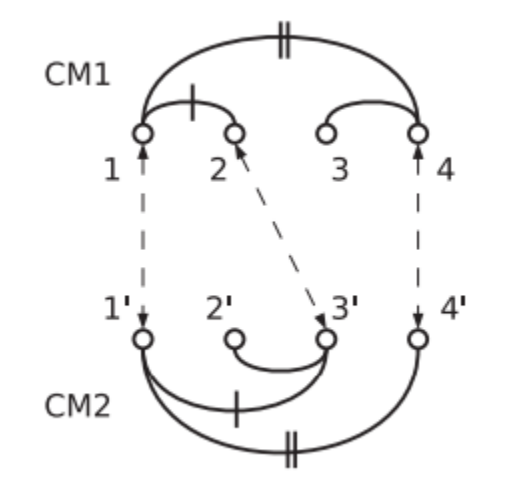
In contact: closer than threshold value
Typically 6 - 16Å



(Pelta et al. 2008)

Alignment with contact maps

- Maximize overlap between two graphs
- Find two subsets of vertices
  Maximize number of *non-crossing* edges



CM1 (1,2,4) matches CM2 (1′, 3′, 4′)
Two edges in common = score 2

How to compute max CMO?

Max CMO is NP hard. Different classes of algorithms

Exact methods

- Optimal solution
- Exponential time (days +)
- Reformat into other graph theory problems

Approximations

- Faster
- Some knowledge about solution quality

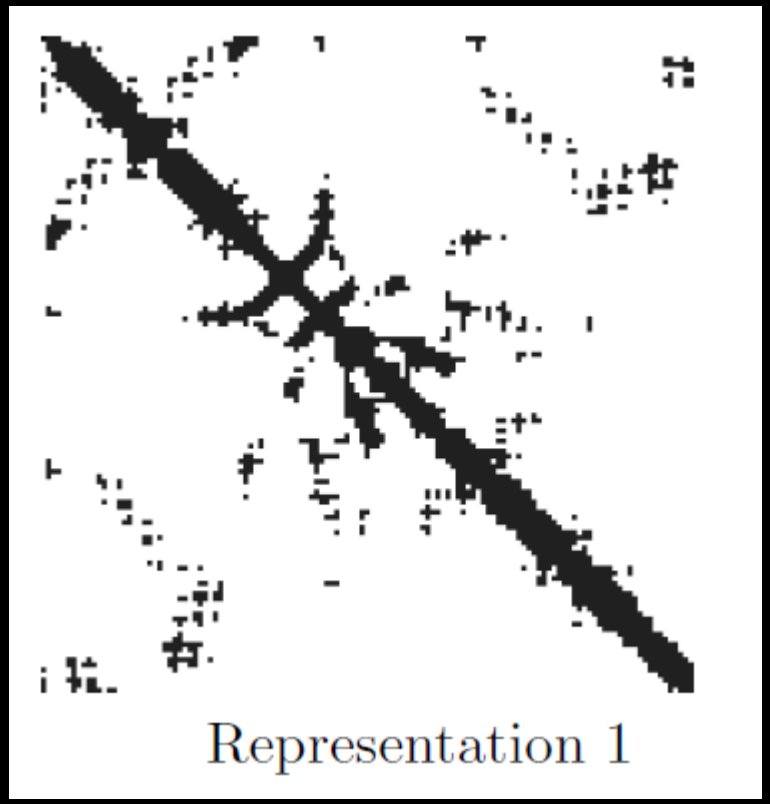
Heuristics

- Fastest
- Good solutions (usually)
- No guarantees about solution quality

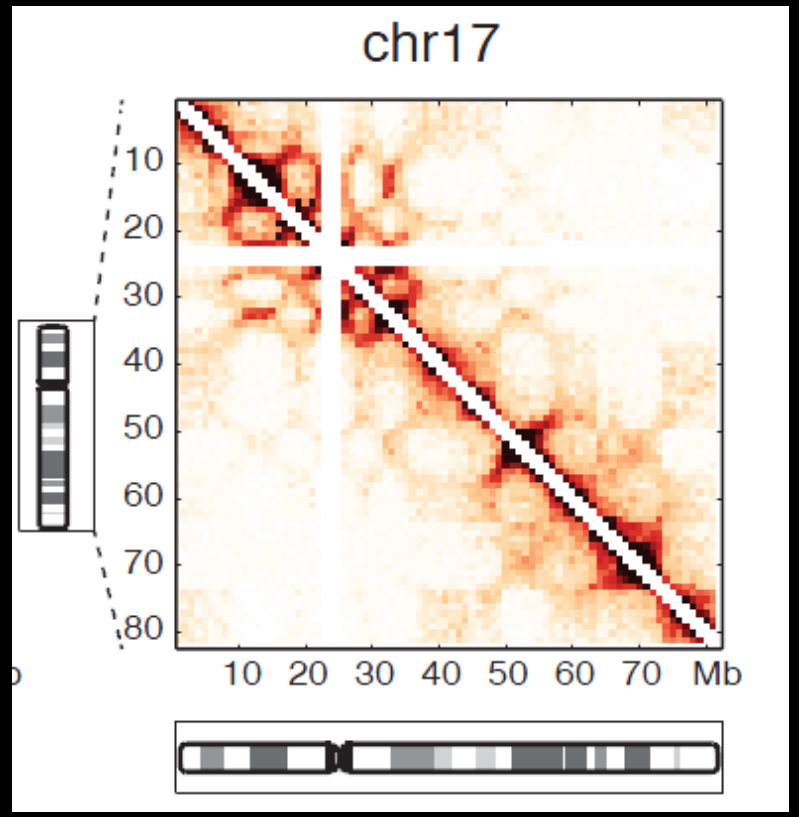
Max CMO applied to Hi-C

Goals
Align chromatin contact maps
Highlight similar regions
- Folding structures
- TAD conservation
- Conservation between
  loci
  chromosomes
  cell lines
  organisms

Problems:
Chromatin contact maps aren't the same as protein maps!
Some assembly required…



Protein contact map
(Pelta et al. 2008)



DNA contact map

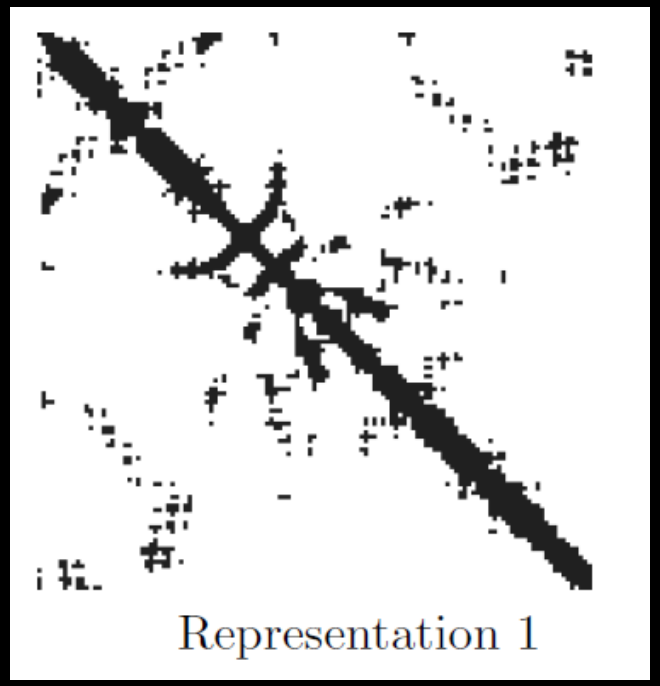
Intuitively similar, but still many differences.
Need to preprocess DNA map to get it in a form applicable to CMO.

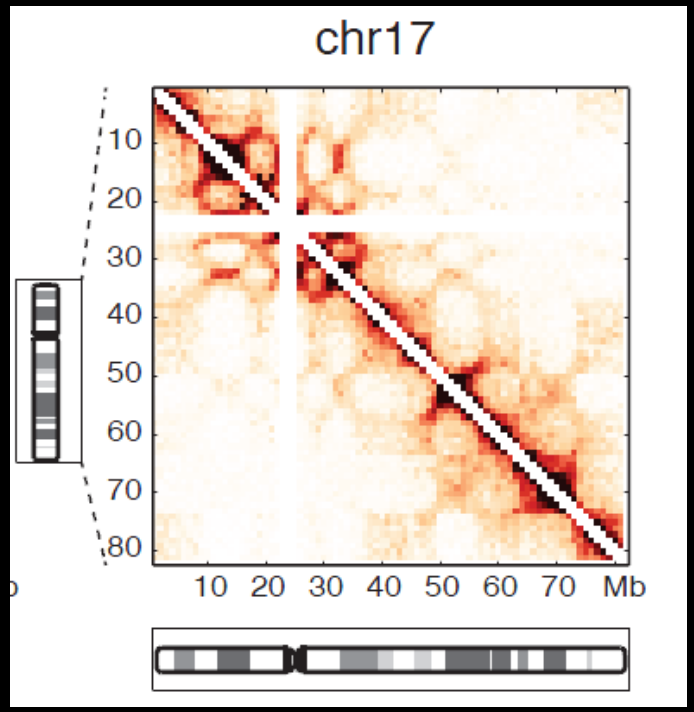
Differences between the two, and some solutions

Protein contact map
- Few contacts
- Values are binary
- Structure is absolute

DNA contact map
- Many contacts
- Values are integers
- Structure is ensemble
- Sequencing errors
- Unsequenced regions

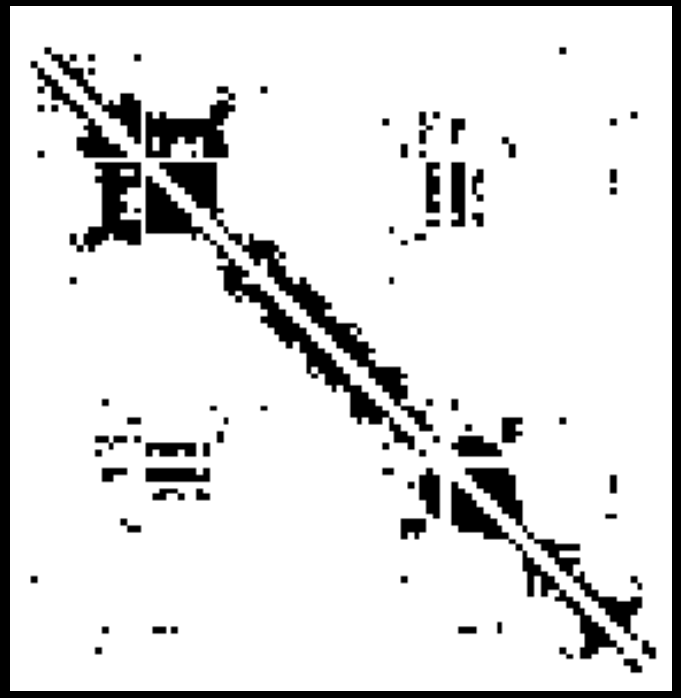
Solutions
- Keep most significant
- Filter at threshold (what?)
- Most significant contacts represent maximum likelihood





Data preprocessing

- Look at 100x100 subset (20Mb chr1)
- Filter some erroneous contacts
- Select 1000 highest (10 contacts per bin)
- Set all values to 1 to get binary



Thoughts?

- Diagonal removed before any data processing done
- Lots of contacts near diagonal
- Few contacts indicating folding
- Not really *most significant*

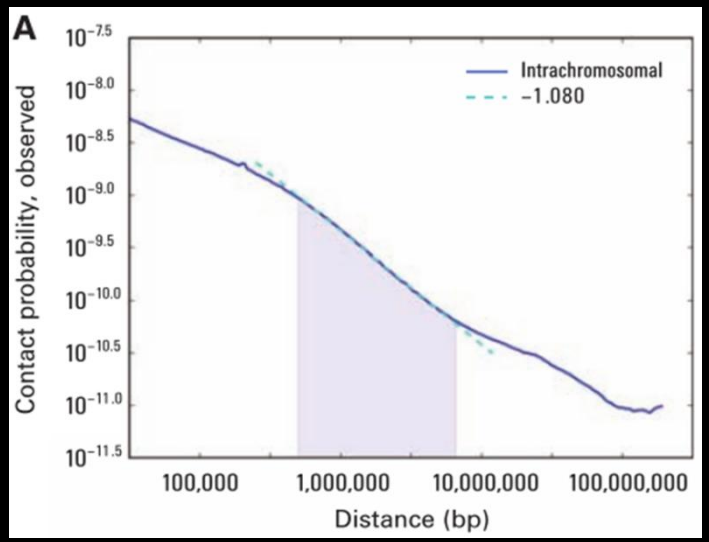
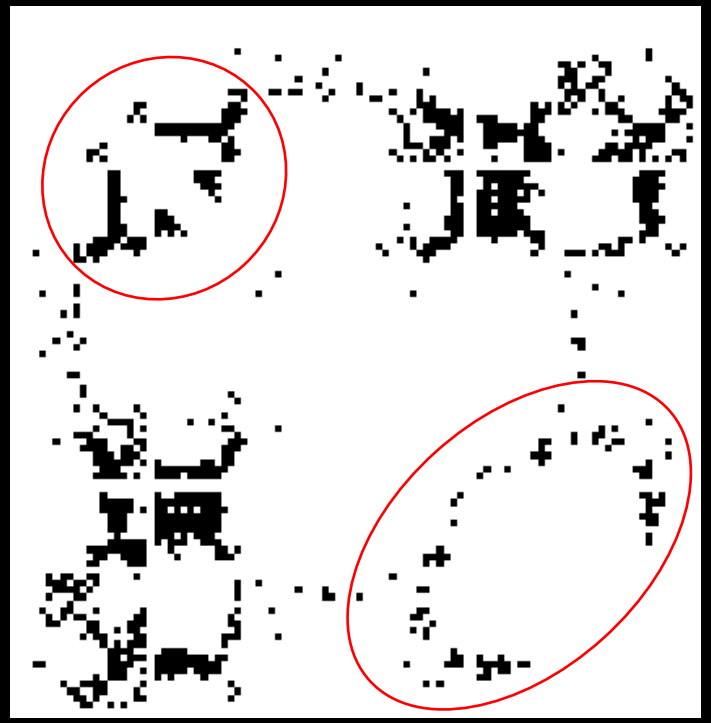


Figure 4A from Lieberman-Aiden et. al 2009
"Contact probability as a function of genomic distance averaged across the genome (blue) shows a power law scaling between 500 kb and 7 Mb (shaded region) with a slope of –1.08 (fit shown in cyan)"

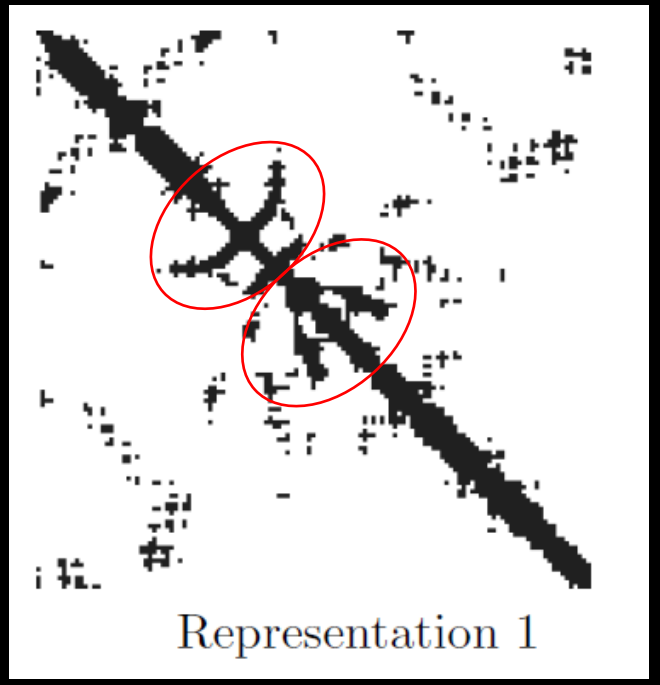
Data preprocessing 2

- Compute observed/expected heatmap
  - Highlights most significant contacts
  - Decrease the amount near diagonal
- Apply same preprocessing as before



Thoughts?

- More significant folding interactions captured
- Some structures like what's seen in protein
- Diagonal??



Results

- Aligned two subsets of the Hi-C matrix
- In practice, 50% of residues overlapped
- 45-55 % overlapped in the alignment
  - Not always the *correct* ones
- An issue with pre processing, I think

Next steps

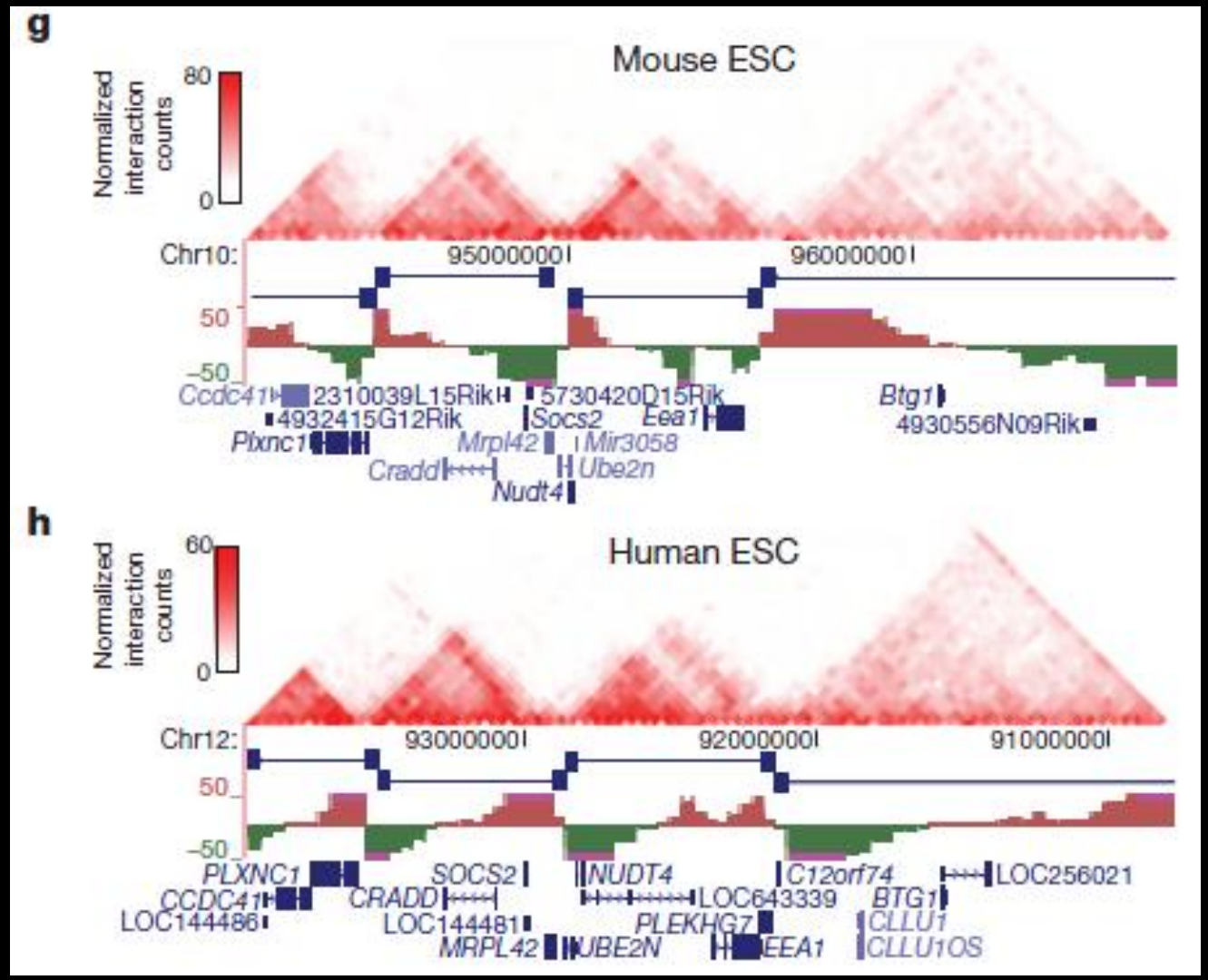
Bing Ren lab at UCSD:

- Mouse Embryonic Stem Cells
- Mouse cortex cells

- Human Embryonic Stem Cells
- Human Fibroblasts

These datasets are not observed over expected, much of the signal is on the diagonal

- Compute O/E and try alignment



How does all this apply to CSCI1820?

- Structure alignment is an analog to sequence alignment
- Sequence alignment can only tell so much
  - In proteins not always relevant
- Algorithms are similar
  - Exponential search space
  - Not as clean shortcuts as sequence alignment

Thank you!

Lancia, G., Carr, R., Walenz, B. & Istrail, S. 101 Optimal PDB Structure Alignments: A Branch-and-cut Algorithm for the Maximum Contact Map Overlap Problem. in *Proceedings of the Fifth Annual International Conference on Computational Biology* 193–202 (ACM, 2001). doi:10.1145/369133.369199

Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19,** 521–532 (2009).

Goldman, D., Istrail, S. & Papadimitriou, C. H. Algorithmic aspects of protein structure similarity. in *40th Annual Symposium on Foundations of Computer Science, 1999* 512–521 (1999). doi:10.1109/SFFCS.1999.814624

Lena, P. D., Fariselli, P., Margara, L., Vassura, M. & Casadio, R. Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics* **26,** 2250–2258 (2010).

Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).
Wit, E. de & Laat, W. de. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26,** 11–24 (2012).

Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Meth* **9,** 999–1003 (2012).

Belton, J.-M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).

Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

Caprara, A., Carr, R., Istrail, S., Lancia, G. & Walenz, B. 1001 Optimal PDB Structure Alignments: Integer Programming Methods for Finding the Maximum Contact Map Overlap. *Journal of Computational Biology* **11,** 27–52 (2004).

Andonov, R., Malod-Dognin, N. & Yanev, N. Maximum Contact Map Overlap Revisited. *Journal of Computational Biology* **18,** 27–41 (2011).

Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).

McCord, R. P. *et al.* Correlated alterations in genome organization, histone methylation, and DNA–lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res.* **23,** 260–269 (2013).

Naumova, N. *et al.* Organization of the Mitotic Chromosome. *Science* **342,** 948–953 (2013).