DNA binding factors

Genome 4D

Bio5488
Ting Wang
3/1/2021

https://www.nature.com/collections/rsxlmsys1k
Protein-DNA Interactions

• Lots of functions:
  • Replication, Repair, Recombination

• Control of expression of the information in the genome

• Specificity of binding sites for regulatory proteins
DNA-binding proteins exhibit a range of specificities

- Restriction enzymes
  - Absolute specificity
  - Discriminate methylations

- Transcription factors
  - Sequence-specific, but
  - Degenerate
  - Range of affinities may be utilized

- Nucleosomes
  - Sequence preferences
  - Translocatable by modeling complexes
Sequence Motifs

- Motif: subsequence with some specific function

- May be in DNA, RNA, protein

- Function may be context dependent
  - Ribosome binding site must be transcribed
  - RNA, protein motifs may depend on structure

- May be gapped or ungapped

- Use model to search for (predict) new sites
  - Models may be simple sequences (regular expressions) or probabilistic patterns

- Modeling approach depends on data available
  - Quantitative/qualitative
Types of Motifs

Motif: Consensus Sequence Pattern
- May include degenerate bases and allow for mismatches
- *Search space is over possible patterns*

Weight Matrix (PWM, Profile, PSSM)
- Might go to higher order models
- *Search space is over possible alignments*
Pattern based algorithms

• Motif length $l$, mismatches $m$; $N$ seqs, $L$ long

• $4^l$ patterns, search for most common (or most significant) allowing up to $m$ mismatches
  • P-value from background distribution
  • Can allow for $m$ mismatches
  • Can allow degenerate positions: $15^l$ patterns
  • Can just search using existing $l$-mers

• Can use suffix tree for efficient search of patterns allowing mismatches
Consensus Sequence Pattern

- Difficult to obtain an optimal consensus for identifying novel sites

- Relative frequency of bases at each positions lost
**Weight Matrix Model**

<table>
<thead>
<tr>
<th>TACGAT</th>
<th>TATAAT</th>
<th>TATAAT</th>
<th>GATACT</th>
<th>TATGAT</th>
<th>TATGTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>-8</td>
<td>10</td>
<td>-1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C:</td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>G:</td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
</tr>
<tr>
<td>T:</td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

- More information than a consensus sequence
- Many ways to determine the weights
- Assumes positional independence
- Requires significant data
Score a site

-24

....A C T A T A A A T G T...

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>T...</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-8</td>
<td>10</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Score a site

43

....A   C   T   A   T   A   A   T   G   T ...

<table>
<thead>
<tr>
<th>A:</th>
<th>-8</th>
<th>10</th>
<th>-1</th>
<th>2</th>
<th>1</th>
<th>-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>C:</td>
<td>-10</td>
<td>-9</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-12</td>
</tr>
<tr>
<td>G:</td>
<td>-7</td>
<td>-9</td>
<td>-1</td>
<td>-1</td>
<td>-4</td>
<td>-9</td>
</tr>
<tr>
<td>T:</td>
<td>10</td>
<td>-6</td>
<td>9</td>
<td>0</td>
<td>-1</td>
<td>11</td>
</tr>
</tbody>
</table>
\[
N(b,i) = \log\left[\frac{F(b,i)}{P(b)}\right]
\]

\[
F(b,i)
\]

\[
S(b,i) = \log[F(b,i)/P(b)]
\]

\[
I(i) = \sum F(b,i)S(b,i)
\]

G. Stormo
## Information Content

### Matrix of Frequencies

<table>
<thead>
<tr>
<th></th>
<th>A:</th>
<th>C:</th>
<th>G:</th>
<th>T:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>G</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.7</td>
<td>0.1</td>
<td>0.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Matrix of Frequencies aka Relative Entropy, Kullbach–Liebler Distance

$$I_{seq} = \sum_{j} \sum_{b} f(b,j) \log_{2} \frac{f(b,j)}{p(b)}$$

Sum is over columns $j$ (the positions), and rows $b$ (the bases)

aka Relative Entropy, Kullbach–Liebler Distance
# Information Content

<table>
<thead>
<tr>
<th>EcoR1</th>
<th>Random</th>
<th>Rap1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAATTC</td>
<td>GCCTAC</td>
<td>TGTATGGGGTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>ACATTC</td>
<td>TGTTCGGGATT</td>
</tr>
<tr>
<td>GAATTC</td>
<td>TCATTC</td>
<td>TGCATGGGGTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>CGACTC</td>
<td>TGTACAGGTG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>GAATTC</td>
<td>TGTATGGATG</td>
</tr>
<tr>
<td>GAATTC</td>
<td>ATATCG</td>
<td>TGTTTCGGGTT</td>
</tr>
<tr>
<td>GAATTC</td>
<td>GAAATG</td>
<td>TGTATGGGGTG</td>
</tr>
</tbody>
</table>
Motif Finding Problem

• A fundamental problem in molecular biology
  • Specific protein and DNA binding
  • Transcription factor binding sites recognition

• Statistical definition:
  • Given some sequences, find over-represented substrings (motif discovery)

• Biological definition:
  • Given some co-regulated promoters, find transcription factor binding model
  • How do we use biology to improve motif finding algorithms?

• Many algorithms/programs developed
  • consensus, gibbs sampling, EM, projection, phylogenetic footprinting, etc.
Motif Finding Algorithms Class I

Single species, multiple genes (planted motif problem)

- random background sequences
- a proper description of a consensus motif $\rightarrow$ better models
- randomly plant copies of the motif into sequences
- define an objective function, and use a search algorithm to find the copies that give a good score
The Data Set: Sequences containing sites for cAMP receptor protein (CRP)

For this case, there are 18 sequences of length 105 bp and we are looking for a motif of width 20 bp. There are 86 different 20 bp subsequences per example and ~7x10^{34} alignments to check.

An (intractable) solution

(Exhaustive algorithm)
Construct every possible combination of alignments and keep the one with the highest information content.

Given a motif of width $w$, and $k$ sequences of length $l$, there are $L = (l-w+1)$ possible locations in each sequence, and $L^k$ alignments to check.
Greedy Algorithm (Consensus)

• Simple version: assume every sequence contains at least one true binding site

• Using each l-mer find best match to generate 2-seq alignments

• Using top K PWMs to search remaining sequences to include a new sequence

• Repeat until all seqs contribute
  • Or objective function is maximized (IC, p-value)
Expectation Maximization (MEME)

- Initial PWM (at random or from average over all potential sites)
- Using current PWM determine probability of all positions being sites
- Re-estimate PWM based on those probabilities
- Continue until convergence – always convergences
- Objective is LLR
Gibbs Sampling

- Similar to EM, but some important differences
- At each iteration pick one site on each seq, chosen by its probability, to update PWM
- Not guaranteed to converge, but tends to increase objective (IC) and plateau
- Can escape local optima
  - Other MCMC algorithms
    - Metropolis
    - Simulated annealing
Gibbs’ Sampling Approach to Motif Discovery

Basic Idea:
- Given “sites”, estimate pattern matrix
- Given “matrix”, pick likely sites according to their probability
- Iterate between those steps until “convergence”

Important details:
- Use “pseudocounts” to avoid prob. = 0
- Sample sites from estimated prob. distrib.
ratios. Equivalently, one may maximize $F$, the sum of the logarithms of these ratios. In the notation developed above, $F$ is given by the formula

$$F = \sum_{i=1}^{w} \sum_{j=1}^{20} c_{i,j} \log \frac{q_{i,j}}{p_j}$$

(2)

where the $c_{i,j}$ and $q_{i,j}$ are calculated from the complete alignment (Fig. 1).

should be supplemented with residue-dependent “pseudocounts” $b_j$ to yield pattern probabilities

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B}$$

(1)

where $B$ is the sum of the $b_j$. The $p_j$ are

Gibbs Sampling

Initialization:
Random assignment of motif locations $a_1 - a_k$

Construct initial matrix $S$ from this alignment
Gibbs Sampling

Update:

Score all possible motif locations in the sequence to be updated with current matrix

Select a new motif placement randomly, but biased toward high scoring regions

Create a new scoring matrix from the new alignment
Gibbs Sampling

How does it end? Eventually you nucleate a few correct placements

The matrix has weak but sufficient scoring power
Gibbs Sampling

An approximately correct matrix rapidly converges, with the subsequent alignments possessing more information content and making better motif window placements.

But notice two suboptimal results: we have one sequence with a placement but no genuine site, and one sequence with two sites but one placement. This is common enough to merit special treatment.
Summary

• The genome encodes much of its own regulation in protein binding sites

• A full description of the regulatory networks will require identifying these sites

• Compact descriptions of the DNA-binding preferences of TFs is afforded by weight matrices

• The information content of an alignment is a measure of specificity

• Weight matrix information for a TF is not enough to rule out false positives

• Multiple experimental techniques exist for identifying sequences harboring binding sites

• A variety of algorithms can be used to identify motifs in unaligned data
Motif Finding Algorithms Class II

Single gene, multiple species (phylogenetic footprinting)

- orthologous background sequences
- sequences linked by a phylogenetic tree
- identify the “best conserved” motif that is under selective pressure
Motif Finding Algorithms Class III

Multiple genes, multiple species

- combination of phylogenetic data and gene regulation
- use phylogenetic data to reduce search space
- use correlation of motif occurrences among orthologous genes to increase signal strength
Protein Binding Microarrays

Custom arrays of 60-mer DNA sequences (~44,000 probes)

Contain all possible 10bp sequences

Each probe contains 27 10-mers

8-mers guaranteed to occur 16 times

ChIP-chip/ChIP-seq

Cross-link protein to DNA

Affinity purify protein-DNA complexes:
- Ab to TF
- Ab to tag on TF
- affinity tag on TF

Reverse cross-links

Identify sequence by hybridization to microarray or by high-throughput sequencing

Bacterial One-hybrid

Genetic selection: survival is dependent on DNA-binding

TF of interest is fused to α-subunit of RNA polymerase

Randomized library of binding sites created and screened for autoactivation

Co-transform with TF and select

Library complexity is limited by transformation efficiency (~10⁹)

High-throughput SELEX

Points of discussion

• Which motif finding algorithm is better?
  • Specific hypothesis
  • Binding site model
  • Search/optimization method and objective function
  • Some basic rules in practice

• Evolution of TF binding sites
  • Binding site turn over
  • Evolution by substitution, in/del, duplication, transposition
  • Co-evolution with TF
  • Impact on shaping regulatory networks

• Motif != binding != function
  • Sensitivity and specificity of wet/dry experiments
  • How to validate?
    • Biological function versus biochemical activity

• Species-specific regulation

• Beyond primary sequence conservation
Challenge of Specificity

• A 7-mer is expected to occur every 16,384 base pairs by chance
• In human, this means $3 \times 10^9 / 16,384 \sim 180,000$ sites in total
• TFBS are usually degenerative
• Total number of genes $\sim 25,000$
• Most of predicted binding sites are false positives!
• Need other restrictive information to reduce false positives
Genome 4D
Outline

• Motivation for studying 3D genome
• Method to investigate 3D genome organization
• 3D genome organization and human diseases
• 4DN project (https://www.4dnucleome.org)
Identify disease or trait-associated variants by genome-wide association studies (GWAS)
Majority of the GWAS hits are located in the non-coding regions

- 5,303 SNPs from 1,675 GWAS studies
- Coding vs. non-coding SNPs
  - Coding SNPs: 7%
  - Non-coding SNPs: 93%

Hnisz et al, Cell, 2013
Majority of the GWAS hits are located in the cis-regulatory elements

Maurano et al. Science 2012
How non-coding mutations contribute to disease

- Gain of TF binding sites;
- Loss of TF binding sites;
- Enhancer hijacking

Onco-gene

Tumor suppressor

How do we find the target genes for distal enhancers?

• Nearest genes by genomic distance;
• Correlation based on gene expression and enhancer activities;
• 3D space proximity;
• CRISPR/Cas9 to delete enhancers and observe their effect on gene expression;
Linking enhancers to target promoters

Gene 1 → Enhancer → Gene 2

Expression → Enhancer activity → Expression

Bone Marrow, Cerebellum, Cortex, Heart, Kidney, Liver, Lung, Mef, ES, Spleen

Correlation coefficient
Regulation at a distance

Preaxial polydactyly

Belgian 2

Enhancer

ZRS
3D genome and transcriptional regulation

Aixel, nature, 2009
Obesity-associated variants within FTO form long-range functional connections with IRX3
Genome organization in mammals

Nucleosomal scale

1 pb - ~10 kb
EPIGENETIC MODIFICATIONS
NUCLEOSOMES

~10 kb - ~800 kb
INTRA-TAD DYNAMICS
TOPologically ASSOCIATING DOMAINS = TADs
CHROMATIN LOOPS

Supranucleosomal scale

~3 Mb
INTER-TAD DYNAMICS
A/B CHROMOSOMAL COMPARTMENTS
NUCLEAR BODY
INTERACTIONS WITH NUCLEAR COMPARTMENTS

Nuclear scale

~100 Mb - ~3000 Mb
NUCLEAR POSITIONING
CHROMOSOME TERRITORIES

Technologies used (and developed) to study genome folding

Can be divided into two broad categories:

1. Imaging
   1. Bright-field
   2. Fluorescence
   3. EM
   4. Fluorescence *in-situ* hybridization (FISH), etc.

2. Genomics
   1. DamID
   2. ChIA-PET
   3. GAM
   4. Chromosome conformation capture-derived, etc.

Adapted from [http://web.uvic.ca/ail/equipment.html](http://web.uvic.ca/ail/equipment.html)
Fluorescence in situ hybridization

- **Cytogenetic technique**

- Uses fluorescent molecules to “paint” (regions of interest) chromosomes in cells often in Metaphase or Interphase

- Aids in analysis of chromosome structure, structural aberrations, ploidy determination, etc.
Multiplexed super-resolution FISH and stochastic optical reconstruction microscopy (STORM)
Capturing Chromosome Conformation

Job Dekker,¹* Karsten Rippe,² Martijn Dekker,³ Nancy Kleckner¹

15 FEBRUARY 2002 VOL 295 SCIENCE www.sciencemag.org
Methodology of “C”-technologies

### 3C
- **Secondary digestion**, circularization (★)
- Inverse PCR
- 3C-qPCR Library

### 4C
- **Secondary digestion, circularization (★)**
- 4C Library: Sequencing

### 5C
- Oligonucleotides: Hybridization, Ligation (★)
- PCR
- 5C Library: Sequencing

### Hi-C
- Biotinylaton
- Ligation
- Sonication
- Purification on Biotin-streptavidin beads
- Ligation of adaptors and PCR
- Hi-C Library: Sequencing

Advantages and drawbacks of “C”-technologies

Table 1. Advantages and limits of 3C-derived methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Genomic Scale Investigated</th>
<th>Advantages</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C-qPCR</td>
<td>~250 kilobases</td>
<td>Very high dynamic range (highly quantitative), easy data analysis</td>
<td>Very low throughput: limited to few viewpoints in a selected region</td>
</tr>
<tr>
<td>4C</td>
<td>Complete genome</td>
<td>Good sensitivity at large separation distances</td>
<td>Genome-wide contact map limited to a unique viewpoint (few viewpoints if multiplex sequencing is used)</td>
</tr>
<tr>
<td>5C</td>
<td>Few megabases</td>
<td>Good dynamic range, complete contact map (all possible viewpoints) of a specific locus</td>
<td>The contact map obtained is limited to a selected region</td>
</tr>
<tr>
<td>Hi-C</td>
<td>Complete genome</td>
<td>Very high throughput (complete contact map)</td>
<td>Poor dynamic range, complex data processing</td>
</tr>
</tbody>
</table>


Hi-C for genome-wide analysis of higher order chromatin structure

Mouse ES cells (from 433 Million Reads)  
**In situ Hi-C**

*In situ Hi-C* maps DNA–DNA contacts occurring in intact nuclei, by proximity ligation.

While initial studies achieved only megabase resolution, the latest study with 15 billion contact reads, reaches kilobase resolution—which is a function of both the size of the restriction enzyme recognition sequence and the sequencing depth of the library.

HiChIP/PLAC-seq

A

Restriction enzyme cutting and biotin fill-in
Proximity ligation

Sonication and immunoprecipitation
DNA purification and biotin enrichment
Pair-end Sequencing

C

chr8:37,470,037-87,571,641 (mm9) 50kb

DHS
H3K27ac
short interactions
short interactions

H3K27ac PLAC-seq

Gene interaction significance
Refseq
Prdx2
Junb
Hook2
Asna1
Tnpo2

Fang et al., 2016
Mumbach et al., 2016
A and B Compartments

Lieberman-Aiden et al., 2009
Hi-C data reveals strong local chromatin interaction domains

\[ DI(x) = \frac{(B - A)}{\text{abs}(B - A)} \left( \frac{(A - E)^2}{E} + \frac{(B - E)^2}{E} \right) \]
The topological domains or TADs
Distribution of contact frequencies – cell to cell variability

Nagano et al. Nature 2013, October
Dynamic chromatin organization during cell cycle

Naumova et al., Science 2013
Overview of features revealed by Hi-C maps

Top: the long-range contact pattern of a locus (left) indicates its nuclear neighborhood (right).

Middle: squares of enhanced contact frequency along the diagonal (left) indicate the presence of small domains of condensed chromatin.

Bottom: peaks in the contact map (left) indicate the presence of loops (right). These loops tend to lie at domain boundaries and bind CTCF in a convergent orientation.

Visualization of Hi-C data

Analysis methods for studying the 3D architecture of the genome. Ferhat Ay and William S. Noble.
Genome Biology (2015)
Test: draw the loop structures
Loop extrusion model

- https://www.youtube.com/watch?v=Tn5qqEqWgW8
One sided extrusion
4D Nucleome Project

a Mapping – molecular genomic mapping of contacts, imaging of contacts and dynamics

b Model building – by constraint modelling and polymer simulation

MMT

ΔT

C Nuclease validation – by genetic and biophysical perturbation experiments

Binding sequences

CRISPR complex

Loop-related proteins

Loop formation

Regions brought into topological contact by polymer rearrangement

Chromosomal regions

Nucleus

Cell

TADs

Nuclear membrane

Loop-related protein

Chromosome
Summary

• Hi-C analysis reveals that the mammalian genome is spatially compartmentalized, and consists of mega-base sized topological domains (also known as TADs).

• Topological domains have been independently observed in flies (Sexton et al. Cell 2012; Hou et al, Mol Cell 2012) and with different approaches (5C, Nora et al., Nature 2012).

• Topological domains are stable across cell types and largely preserved during evolution, suggesting that they are a basic property of the chromosome architecture.

• Partitioning of the genome into topological domains would naturally restrict the enhancers to selective promoters.

• Long-range looping interactions between enhancers and promoters correlate with higher transcriptional responsiveness of promoters.

• Cell specific enhancer/promoter interactions are formed in each cell type, some time prior to activation of the genes, and are not significantly altered by transient signaling induction.

• Pre-existing, lineage specific chromatin looping interactions between enhancers and promoters predict transcriptional responses to extracellular signaling, suggesting that chromatin conformation is another layer of transcriptional control.

• Hi-C types of assays interrogate population average!