Pattern Recognition in Genomics
DNA Motifs and Human Gene Editing (i.e. CRISPR-Cas9 Off-targets)

Ka-Chun Wong, Jiecong Lin
Dr. WONG Ka Chun (黃家駿博士)
Assistant Professor
(PhD Supervisor)

Ka-Chun has spent 3.5 years (2012-13 departmental average: 8 years after master degree) to finish his PhD degree in Department of Computer Science at University of Toronto in 2014. He is merited as the first associate editor outside USA and Germany for the open-access and open-peer-review journal, BitData Mining, in 2016. He is also on the editorial board of Applied Soft Computing since 2016. He was invited as the plenary/keynote speaker for ICBR 2017 and ISACIT 2018. He is one of the few CityU professors who are invited to deliver a speech in the HKU & ND workshop in 2016. In addition, he has solely edited 2 books published by Springer and CRC Press, attracting 30 peer-reviewed book chapters around the world (i.e. Argentina, Australia, Belgium, Brazil, China, Egypt, France, Germany, Hong Kong, India, Japan, Spain, USA). In 2017, he has solely authored a journal paper about Motility in published on Bioinformatics, demonstrating a solid example for his PhD students at CityU. Lastly, Ka-Chun explicitly showed his track records since the 2003 SARS outbreak below, encouraging all peer students under the Hong Kong education system to trust and embrace their dreams under diverse settings.

Award and Achievement

- 2018 “General Research Fund” (57 out of 175 CS professors), RGC, Hong Kong.
- 2017 “General Research Fund” (56 out of 188 CS professors), RGC, Hong Kong.
- 2016 “Early Career Scheme” (Largest 2015 ECS in CS area), RGC, Hong Kong.
- 2015 “Assistant Professorship” CS, City University of Hong Kong.
- 2015 “Doctor of Philosophy” CS, University of Toronto.
- 2013 “Acres - Joseph Yoon Memorial Fellowship” CS, University of Toronto.
- 2012 “Kwok Sau Po Scholarship” SGS, University of Toronto.

Research Interests

- Bioinformatics
- Applied Machine Learning
- Applied Data Mining
- Natural Computing
- Computational Biology
- Computational Science
- Interdisciplinary Research
Research Collaboration on ChIP-seq

Name: YU Jun 于君
Title: Professor
Home Department: Medicine and Therapeutics
Institute of Digestive Disease
Email: junyu@cuhk.edu.hk
Office Phone No.: (852) 3753 6099
Office Fax No.: (852) 2144 5330
Office Address: Room 707, Li Ka Shing Medical Sciences Building, Prince of Wales Hospital
Research Collaboration on Bioinformatics

Yue Li
(2015-2018) Postdoctoral associate, Computational Biology, CSAIL, MIT

I will be an Assistant Professor for School of Computer Science at McGill University in January 1 2019!
Welcome to visit my lab!

I'm a postdoctoral associate from Prof. Manolis Kellis research group at Computer Science and Artificial Intelligence Laboratory (CSAIL) at Massachusetts Institute of Technology. My postdoc research is mainly interested in developing machine-learning methods and bioinformatics tools to reveal meaningful patterns involving genetics, epigenetics, expression dynamics that are together associated with complex human diseases.
Research Collaboration on Biomedical Text Mining
Outline

• Background

• Pattern Recognition on DNA Motifs:
  – DNA Motifs on Protein Binding Microarray [1]
  – DNA Motifs on Chromatin Interactions [2]

• Pattern Recognition on CRISPR-Cas9 Off-targets [3]

References


Background
Background

Central Dogma of Molecular Biology: “DNA makes RNA, RNA makes Protein”

Background

Protein-DNA Binding

• The binding between proteins (e.g. Transcription Factors, TFs) and DNA (e.g. Transcription Factor Binding Sites, TFBSs) play an important role.

• TFs bind in a sequence-specific manner to TFBSs to regulate gene transcription.
Background

• To fully understand a gene’s function, it is essential to identify the TFs that regulate the gene and the corresponding TF-binding sites (also known as DNA Motifs).

• DNA motifs are relatively short (10–20 bp) and highly degenerate sequence motifs, which make their effective identification a challenging task.

A motif logo example
DNA Motifs on Protein Binding Microarray [1]

DNA Motifs on Protein Binding Microarray [1]

- A fundamental bottleneck in DNA motif identification is the lack of quantitative binding affinity data for a large proportion of the TFs.

- The advancement of new high-throughput technologies such as ChIP-chip, ChIP-seq, has made it possible to determine the binding affinity of these TFs. However, Chip-chip and Chip-seq are in vivo (“within the living”) approach with unpredictable throughput. Co-factors can also obscure the results.

- Therefore, an in vitro high-throughput approach, known as Protein Binding Microarray (PBM) is introduced. PBM enables us to measure the DNA sequence binding of TFs in vitro (test-tube) to all the possible k-mers (DNA short sequence with length = k).

Example (k = 8)
CCATGGGC
ATGCGGGA
CATGGGTC
**DNA Motifs on Protein Binding Microarray** [1]

<table>
<thead>
<tr>
<th>Segments of a DeBruijn Sequence</th>
<th>Normalized Signal Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAAAAAAACGGGTCTGTGTGTGTGTTTCCATGCGCAGCTGTGTTCCGTTGTCGCTGTG</td>
<td>1421.285654</td>
</tr>
<tr>
<td>CAGGGAAAGCAGGTTTTCCATTGAAATCAACCCATCAGATCGTTGTCGCTGTG</td>
<td>719.1041392</td>
</tr>
<tr>
<td>ACCATCGGTGTGTCATACGCCCCCATTACGACGTTGTGTCGCTGTG</td>
<td>1762.812037</td>
</tr>
<tr>
<td>ATATCGAGCAAGTAAAAACGCGAAAAATCTTTTGGAGCTGTGTTCCGTTGTCGCTGTG</td>
<td>465.8865101</td>
</tr>
<tr>
<td>ATTTGGAATTATATGGCCAAGATAGTTAGAATACCTGTGTCGCTGTG</td>
<td>588.0584476</td>
</tr>
<tr>
<td>GAAATCCCTACCAAATGTATGTAAGACTTGGTGATGTCGCTGTG</td>
<td>582.1760333</td>
</tr>
<tr>
<td>CTTTCGTATACAAAGTATCCCGTACTACGCCTTGGGCTCGTTGTCGCTGTG</td>
<td>710.6338075</td>
</tr>
<tr>
<td>CCGGTGTGGTATCCTGGTGCTATGGCTGTG</td>
<td>934.8380349</td>
</tr>
<tr>
<td>GCTAGATTCCCTACGCCATCGTCAGACAGAAGCAGCTGTG</td>
<td>741.7662929</td>
</tr>
<tr>
<td>ACAGCAACGGGAGCCATGAGCAGGACGCTACATAATGCTGTG</td>
<td>18507.18435</td>
</tr>
<tr>
<td>ACATACTGTTCAGGTAAGCCATAGCTCTTCCAGGTCGTCG</td>
<td>759.8673106</td>
</tr>
<tr>
<td>ATTTCCAGTCTCCTGCCTGGTGATGTCGCTGTG</td>
<td>868.2056512</td>
</tr>
<tr>
<td>TGCTTTGCAATTACGAACATAAAGTTGCTCCTTGTGTCG</td>
<td>597.3772212</td>
</tr>
<tr>
<td>TCCCTTGCTCCCGTCTCAGCAACTGAAACATAGCATGTCTGTG</td>
<td>1156.546346</td>
</tr>
<tr>
<td>CATAAGCATTATCCGATGTCGTCGAGAAGCTTTATTGTTGTCG</td>
<td>982.9229197</td>
</tr>
<tr>
<td>TCCTTTAGGCGATTCTCCAGCAGTGGTCGTCGCTGTG</td>
<td>491.8523141</td>
</tr>
<tr>
<td>GCTGGCAATTACCAGCCGGCGGAAATAGTCCGGCGGCTGTGTTCCGTTGTCGCTGTG</td>
<td>673.6856787</td>
</tr>
</tbody>
</table>
Motivation

• Given PBM data, a new branch of algorithms are needed to take into account the quantitative affinity data to uncover a motif model, i.e. motif discovery in PBM data.

• Existing algorithms including MatrixREDUCE, MDScan, PREGO, RankMotif++, and Seed and Wobble all make use of the most common motif model, the Position Weight Matrix (PWM).

Source:
http://en.wikipedia.org/wiki/Position_weight_matrix
Motivation

• However, the adjacent nucleotide independence assumption that PWM makes is **unrealistic** in many cases [a].

• Although a recent attempt has been made to generalize PWM, the insertion and deletion operations between adjacent nucleotide positions are **still challenging** [b].

---


Contribution

• In this work, we developed a hidden Markov models (HMM)-based approach (kmerHMM) to model the dependence between adjacent nucleotide positions rigorously, which outperforms existing approaches.

• We also discovered that HMM can be used to deduce multiple binding modes for a given TF, which is verified by comparison with existing algorithms and wet-lab literature evidence.
kmerHMM

• For each dataset, we train a Hidden Markov Model (HMM) based on DNA k-mers with binding intensities obtained from Protein Binding Microarray (PBM) data to represent the binding specificity of the TF.

• We show that the trained HMM can be used to
  – Predict if a DNA sequence can be bound by each TF
  – Rank the binding intensities of the DNA sequences
  – Identify the multiple binding modes for each TF
Methodology (kmerHMM)

Step 0 - 3: Training (on one array)

Step 4a: Testing on another array

Step 4b: Analyzing the trained HMM

Figure 1. An HMM approach for multimodal motif discovery from PBM data. (1) Positive (bound) k-mers are selected from the training DNA probe sequences (e.g. Array #1). (2) The positive k-mers are aligned using a multiple sequence alignment method. (3) The aligned positive k-mers are input for training an HMM using Baum-Welch training in an unsupervised fashion. (4a) The trained HMM is tested on the testing DNA probe sequences (e.g. Array #2). (4b) The trained HMM can be analyzed and visualized using N-Max-Product algorithm.
Step 0: Identify all distinct k-mers

- For each DNA sequence, trim the primer region.

- For each DNA sequence, use a sliding window of length k to identify all the distinct k-mers (with considerations on their reverse complements)

- k = 8 is used throughout the study, in order to compare with the other approaches

Example (k = 8)
CCATGGGC
ATGCCGGA
CATGGGTC
Step 1: Build a Signal Intensity List and Associate the median to the k-mer

- For each distinct DNA k-mer, build a signal intensity list.
- For each DNA k-mer, use the median of the signal intensity list to represent the signal of the k-mer.
- Pick the top k-mers based on their signals. (For example, above median plus several median absolute deviations)

\[ m_y > m_i + 4\sigma \]

<table>
<thead>
<tr>
<th>DNA kmer</th>
<th>Signal Intensity List</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCATGGGGC</td>
<td>{580, 601, 620}</td>
<td>601</td>
</tr>
<tr>
<td>CATGGGCA</td>
<td>{579, 601}</td>
<td>590</td>
</tr>
<tr>
<td>ATGGGCAA</td>
<td>{400, 579, 580}</td>
<td>579</td>
</tr>
<tr>
<td>ATGCCGGA</td>
<td>{500, 552, 600}</td>
<td>552</td>
</tr>
<tr>
<td>TGCCGGAC</td>
<td>{500, 552, 700}</td>
<td>552</td>
</tr>
</tbody>
</table>

* The example has been simplified for presentation purpose.
Step 2: Align the k-mers using Multiple Sequence Alignment (MSA)

- Select only the top k-mers
- Align the top k-mers using multiple sequence alignment.

Example
CCATGGGC
GGGCATTT
CGGATTTT
GGACTTTA
TTTACCAT

Multiple Sequence Alignment (MSA)

```
  - - C C A T G G G C - - -
  G G G C A T T T - - - -
  - C G G A T T T T C - - - -
  - G G A C T T T T A - - - -
  - - - - - - T T T A C C A T
```

* The example has been simplified for presentation purpose.
Step 3: Train a HMM based on the Multiple Sequence Alignment

- **Train a HMM** based on the Multiple Sequence Alignment as input, using the Baum-Welch Training algorithm.

- 50 hidden states are used.

- As each HMM is initialized randomly, the training is repeated for 10 times to avoid any suboptimal convergence.

* The example has been simplified for presentation purpose.
Step 4a: Verifications on Testing DNA sequences

• Test the HMM on another array, i.e. if the HMM is trained on array #1, it will be tested on array #2 (and vice versa).

• Two testing procedures:
  1. Given the testing DNA sequences only, ask the algorithm to rank them and compare to original ranking.
  2. Given the testing DNA sequences only, ask the algorithm to classify which one is likely to be bound by the TF (+ve class) and which one is less likely to be bound by the TF (-ve class).
Step 4b: Find Multimodal binding

- Given a HMM, we find the Top N most probable state transition paths for a visit of L-states using the N-Max-Product Algorithm, i.e. the generalized version of the Viterbi algorithm, where L is the alignment length of the k-mers.

- For each probable path, we can map a position-specific weight matrix (PWM) based on the emission probabilities in each state.

- Single-linkage Hierarchical clustering can be applied on these PWMs to identify multiple binding modes of the TF.
Two Datasets

• **Dataset 1**, 5 TFs
  - Cbf1
  - Ceh-22
  - Oct-1
  - Rap1
  - Zif628

• Each TF contains 2 arrays.
• These PBM data have also been processed by algorithms including MatrixREDUCE, MDScan, PREGO, RankMotif++ and Seed and Wobble.

• **Dataset 2**, 42 TFs
  - Arid3a
  - Ascl2
  - Bcl6b
  - Bhlhb2
  - E2F2
  - ...
  - ...

• Each TF contains 2 arrays.
• These PBM data have also been processed by the algorithm RankMotif++.
Protein Binding Microarray Data
Two arrays (data replicates) for each TF

- **Array #1 for TF Arid3a**

- **Array #2 for TF Arid3a**

  >= 40,000 sequences, Each sequence has a length of 35 ~ 40 bp
Results Overview

• Step 4a: Verifications on Testing DNA sequences
  – **Experiment 1**: to evaluate the algorithms by ranking and classification on dataset 1, which consists of 5 TFs.
  – **Experiment 2**: to evaluate the algorithms by ranking and classification on dataset 2, which consists of 42 TFs.

• Step 4b: Find multimodal binding
  – **Experiment 3**: to analyze the HMM trained on dataset 1
  – **Experiment 4**: to analyze the HMM trained on dataset 2
Experiment 1 (Ranking)

- The algorithms are trained on array #1 and tested on array #2 (or vice versa).

- On the testing dataset, given only the DNA sequences, we ask the competing algorithms to rank them for binding preference.

- The similarity between the predicted rank and the actual rank can be measured by Spearman’s rank correlation coefficient. The higher the value is, the better the performance.

\[
\rho = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]

, where \(x_i, y_i\) are ranks.
HMM: Use Forward Algorithm to Predict the binding intensity

The HMM approach
1. Use a sliding window to scan all possible k-mers on a sequence
2. Use **forward algorithm** to associate each k-mer with an occurring probability
3. Select the k-mer with the maximum probability value and assign this value to the whole sequence as the predicted binding intensity.

Forward algorithm: Given a HMM, and a k-mer, the algorithm computes how likely is the k-mer generated by the HMM.

Quantitative measure for ranking. Mathematically, given a DNA sequence $D = d_1d_2...d_T$, we compute its predicted binding preference $B(D)$ as:

$$B(D) = \max_{\theta} P(d_1d_2d_3...d_T; \theta)$$

where $P(d_1d_2d_3...d_T; \theta)$ can be computed using the forward algorithm (65), similar to the training procedure described in the previous section.
From those results, we can observe that kmerHMM performs better than other methods on three datasets (Cbf1, Oct-1 and Zif238). (except 8-mer gold standard)

On the two other data sets (Ceh-22 and Rap1), kmerHMM is not the top performer but is close.

In the case of Rap1, kmerHMM performed slightly worse than other methods. The consensus binding motif for Rap1 is 13 nt long. kmeHMM only considers motifs of 8 bp; therefore, it is at an disadvantage.
Experiment 1 (Classification)

- On the testing dataset, given only the DNA sequences, ask the algorithms to classify them by positive binding (+) or negative binding (-).

- The algorithms are trained on array #1 and tested on array #2 (or vice versa).

- The problem can be considered as a binary classification problem and thus can be evaluated by AUC.

How to obtain the actual class labels?
Given a list of actual intensities, compute the median \( m \) and the standard deviation \( \sigma \). A sequence is +ve if its intensity \( \geq m + n\sigma \), otherwise, -ve.

How to obtain the predicted class labels?
After using forward algorithm, each sequence is assigned a predicted binding intensity. We find a threshold such that specificity = 99%, and use this threshold to perform the classification.
Results of Experiment 1 (Classification)

kmerHMM is the best comparing to other algorithms, except on Ceh-22. (except the 8-mer gold standard)
Comparing with RankMotif++, which is also a k-mer approach.
Results of Experiment 3: Find multimodal binding in dataset 1

- Run N-Max Product Algorithm on the HMM trained for each TF in dataset 1 to find the top N most probable paths.

- Analogy
  - Viterbi algorithm: Given a HMM and an observed sequence of length L, compute the most probable path of state transitions of Length L.
  - N-Max Product Algorithm: Given a HMM and a parameter L, compute the top N most probable path of state transitions of Length L. (In the study, N is set to a number until the probability of the N+1-th path is < 0.001.)

Each path can be used to generate a PWM, by simply use the emission probability as a matrix column.
Top N most probable paths

- A single-linkage hierarchical clustering is applied on the PWMs generated by the Top N most probable paths.

- A dendrogram cutoff was chosen such that the mean of a cluster validity measure, i.e. the mean of silhouette values (the higher, the better), was the highest.

- Oct-1 was chosen as an illustrative example such that it was found to have two clusters.
This experiment also reflects that HMM modeling is necessary for multimodal motifs, comparing with other modeling in which state transition path topology is restricted to a principal state transition path manually.
Discussion

Novelty lies in two aspects.

1. kmerHMM outperforms existing methods by using HMM to derive a model to represent PBM data. To our knowledge, this is the first instance that HMM is used in representing PBM data.

2. Second, kmerHMM incorporates N-max algorithm and can derive multiple motif matrix models to represent PBM data.

To the authors’ knowledge, this work is the first study incorporating HMMs into the PBM motif discovery problem.

In a broader sense, this work is also the first study incorporating max-product algorithms into the general motif discovery problem explicitly.
Discussion

Limitation: The use of sliding window

- The potential drawback of the proposed approach is that it relies on a sliding window to segment DNA probe sequences into individual k-mers, which may lose the sequence context information.

- We expect such a limitation will be alleviated when a future improved PBM technology can generate binding affinity for longer k-mers (i.e. higher k value).
Discussion

Implication: Not limited to motif discovery

• From the research framework, we can learn how to make use of the k-mers to train a HMM for pattern recognition, rather than a simple PWM.

• From the state transition path analysis, we can observe that HMM training is effective in handling multimodal pattern recognitions, which other modeling methods may not be able to handle.
In this study, we proposed a novel computational pipeline for PBM motif discovery in which HMMs are trained to model the nucleotide dependency of DNA motifs, and Belief Propagation is used to elucidate multiple motif models from each trained HMM.

The new algorithm, kmerHMM, is compared with other existing methods on benchmark PBM data sets and demonstrated its effectiveness and uniqueness.

We also demonstrated that kmerHMM can capture multiple binding modes of a DNA-binding protein, for which a single position weight matrix (PWM) model is unable to do.

We foresee that a method like kmerHMM will provide biological insights and will be useful in this arena or other domains.

DNA Motifs on Protein Binding Microarray  [1]

DNA Motifs on Chromatin Interactions [2]

DNA Motifs on Chromatin Interactions [2]

https://www.dovepress.com/cr_data/article_fulltext/s3500-35598/img/fig2.jpg
Motivation

• A **topologically associating domain (TAD)** is a self-interacting genomic region, meaning that DNA sequences within a TAD physically interact with each other more frequently than with sequences outside the TAD. These three-dimensional chromosome structures are present in animals as well as some plants, fungi, and bacteria. TADs can range in size from thousands to millions of DNA bases.

• The functions of TADs are not fully understood, but in some cases, disrupting TADs leads to disease because changing the 3D organization of the chromosome disrupts gene regulation. The mechanisms underlying TAD formation are also complex and not yet fully elucidated, though a number of protein complexes and DNA elements are associated with TAD boundaries.

https://en.wikipedia.org/wiki/Topologically_associating_domain
Motivation

• However, the underlying long-range chromatin interactions between different DNA motifs, which are related to various diseases (pmid24496500), are still under-explored due to the low-resolution nature (5-10kbp) of the existing studies (pmid24141950).

• Other studies are limited to few proteins such as CTCC-binding factor (CTCF) and the cohesin complex. Briefly, CTCF is the well-known TF which can control chromatin interactions for gene regulation (pmid24614316, pmid26686651) while the cohesion complex consists of four proteins (e.g. SMC1, SMC3, alpha-kleisin, and STAG1-3 in human) which can collectively regulate the modular shapes of chromatin interactions (pmid24335803).

• Therefore, it is very important to develop de novo algorithms for discovering DNA motif pairs on those chromatin interactions where the de novo nature is necessary for the vast variety of rapidly growing sequencing data in different cell types.
Motivation

• Unfortunately, the pairing setting of the chromatin interactions imposes difficulties in applying the existing de novo cis-regulatory module discovery which input is limited to unpaired sequences (pmid15297614).

• Therefore, MotifHyades is proposed for de novo DNA motif pair discovery on paired sequences (i.e. long-range chromatin interacting promoter-enhancer pairs) as an "one-stop" model with linear complexity in this study.
Proposed Approach (MotifHyades)
Proposed Approach (MotifHyades)

Given a set of $T$ paired sequences (e.g. promoter-enhancer sequence pairs on long-range chromatin interactions) $D = \{ (PSeq^{(t)}, ESeq^{(t)}) | t \in \mathbb{N}, t \leq T \}$, our objective is to find $K$ DNA motif pairs $M = \{ (M^i_P, M^i_E) | i \in \mathbb{N}, i \leq K \}$ with prior occurrence probabilities $P = \{ \pi^i | i \in \mathbb{N}, i \leq K \}$ that maximize the likelihood:

$$
\arg \max_{\Theta} L = \prod_{t=1}^{T} \sum_{x^{(t)} = 1}^{K} P(PSeq^{(t)}, ESeq^{(t)} | x^{(t)}; \Theta) P(x^{(t)}; \Theta)
$$

where a MotifHyades model can be defined as $\Theta = (M, P)$. Semantically, a hidden variable $x^{(t)}$ is defined to denote the index of motif pair; each sequence pair $(PSeq^{(t)}, ESeq^{(t)})$ probabilistically harbor the $i$-th motif pair $(M^i_P, M^i_E)$ with the prior probability $\pi^i = P(x^{(t)} = i)$. 
Proposed Approach (MotifHyades)

However, the sequence locations of the motif pair instances are not known in advance. Therefore, the hidden variables $z_{i_P}^{(t)}$ and $z_{i_E}^{(t)}$ are defined to denote the sequence locations of each motif pair instance ($M_P^i$, $M_E^i$). Therefore, the joint probability for the $t$-th sequence pair occurrence $P(\text{PSeq}^{(t)}, \text{ESeq}^{(t)}|x^{(t)}; \Theta)$ is modeled as:

$$
\sum_{z_{i_P}^{(t)}=1}^{L_P^{(t)}} \sum_{z_{i_E}^{(t)}=1}^{L_E^{(t)}} P(\text{PSeq}^{(t)}, \text{ESeq}^{(t)}, z_{i_P}^{(t)}, z_{i_E}^{(t)}|x^{(t)} = i; \Theta)
$$

where each possible sequence positions are modeled by the hidden variables $z_{i_P}^{(t)}$ and $z_{i_E}^{(t)}$ for the $(t)$-th sequence pair $(\text{PSeq}^{(t)}, \text{ESeq}^{(t)})$. On the other hand, conditional independence is assumed within each sequence pair $(\text{PSeq}^{(t)}, \text{ESeq}^{(t)})$, given $x^{(t)} = i$, such that:

$$
P(\text{PSeq}^{(t)}, \text{ESeq}^{(t)}, z_{i_P}^{(t)}, z_{i_E}^{(t)}|x^{(t)} = i; \Theta)
$$

$$
= P(\text{PSeq}^{(t)}, z_{i_P}^{(t)}|x^{(t)} = i; \Theta) P(\text{ESeq}^{(t)}, z_{i_E}^{(t)}|x^{(t)} = i; \Theta)
$$
Proposed Approach (MotifHyades)

At this point, different motif modeling strategies can be adopted here. However, a recent study indicates that the position frequency matrix modeling is simple but robust enough for DNA motif modeling (Weirauch et al., 2013). Therefore, the following formula is defined for $P(PSeq^{(t)}, z_{iP}^{(t)}|x^{(t)} = i; \Theta)$:

$$\frac{\prod_{l=1}^{L} \prod_{i=1}^{S_{l}} \frac{M_{P}(i, PSeq^{(t)}[s + z_{iP}^{(t)} - 1])}{BG(PSeq^{(t)}[s + z_{iP}^{(t)} - 1])}}{\prod_{l=1}^{L} \prod_{s=1}^{S_{l}} \frac{M_{P}(s, PSeq^{(t)}[s + z_{iP}^{(t)} - 1])}{BG(PSeq^{(t)}[s + z_{iP}^{(t)} - 1])}}$$

Similarly, $P(ESeq^{(t)}, z_{iE}^{(t)}|x^{(t)} = i; \Theta)$ can be derived as follows:

$$\frac{\prod_{l=1}^{L} \prod_{i=1}^{S_{l}} \frac{M_{E}(i, ESeq^{(t)}[s + z_{iE}^{(t)} - 1])}{BG(ESeq^{(t)}[s + z_{iE}^{(t)} - 1])}}{\prod_{l=1}^{L} \prod_{s=1}^{S_{l}} \frac{M_{E}(s, ESeq^{(t)}[s + z_{iE}^{(t)} - 1])}{BG(ESeq^{(t)}[s + z_{iE}^{(t)} - 1])}}$$

where, in this study, we have adopted the simplest sequence background model in which $BG(c)$ denotes the $c$-th nucleotide frequency of the background sequences (i.e. Markov model of zeroth order) to ensure that the background model complexity can be comparable to the position frequency matrix modeling for effective pattern recognition (Weirauch et al., 2013). On the other hand, $M_{P}(j, c)$ and $M_{E}(j, c)$ denote the $c$-th nucleotide frequency at the $j$-th sequence position of the DNA motif matrix models $M_{P}$ and $M_{E}$ respectively.
Proposed Approach (MotifHyades)

3.1 MotifHyades(EM)

By taking partial derivatives to the expected complete data likelihood $E[\log L]$ (plus adding Lagrange multipliers to the sum-to-one constraints) with respect to parameters $\Theta$ to zero, the expectation maximization method is derived for model training, named as MotifHyades(EM). Details can be found in the supplementary data.

3.2 MotifHyades(Gibbs)

Although the previous model training algorithm is the exact form of expectation maximization, it is noted that the overall computational complexities are proportional to the total sequence lengths which can scale and summed up to millions. Therefore, a variant is derived with the partial help of Gibbs Sampling, named as MotifHyades(Gibbs), as elaborated in the supplementary data.
MotifHyades (EM)

**E-Step:**

By taking partial derivatives to the expected complete data likelihood $E[\log L]$ (plus adding Lagrange multipliers to the sum-to-one constraints) with respect to parameters $\Theta$ to zero, the expectation maximization method is derived for model training, named as MotifHyades (EM).

$$
E\text{-step:}
$$

$$
P(x^{(i)} = t | D; \Theta) = \frac{P(PSeq^{(i)}, ESeq^{(i)} | x^{(i)} = t, \Theta) P(x^{(i)} = t | \Theta)}{\sum_{t} P(PSeq^{(i)}, ESeq^{(i)} | x^{(i)} = t, \Theta) P(x^{(i)} = t | \Theta)}
$$

where $F_p^{(i)}(i, j, c)$ denotes the expected occurring frequency of the $c$-th nucleotide at the $j$-th sequence position of the DNA motif model $M_p$ while considering all possible positions of the $t$-th sequence $PSeq^{(i)}$. Similarly, $F_E^{(i)}(i, j, c)$ denotes the expected occurring frequency of the $c$-th nucleotide at the $j$-th sequence position of the DNA motif model $M_E$ while considering all possible positions of the $t$-th sequence $ESeq^{(i)}$.

**M-Step:**

$$
M\text{-step:}
$$

where $\pi_t = \frac{\sum_{j=1}^{T} P(x^{(i)} = t | D; \Theta)}{T}$

$$
M_p^{(j)}(j, c) = \frac{\sum_{t=1}^{T} P(x^{(i)} = t | D; \Theta) F_p^{(i)}(i, j, c)}{\sum_{t=1}^{T} P(x^{(i)} = t | D; \Theta)}
$$

$$
M_E^{(j)}(j, c) = \frac{\sum_{t=1}^{T} P(x^{(i)} = t | D; \Theta) F_E^{(i)}(i, j, c)}{\sum_{t=1}^{T} P(x^{(i)} = t | D; \Theta)}
$$

Briefly, all the model parameters $\Theta = (M, P)$ are randomly initialized at the beginning. E-step and M-step are then alternated and repeated until the change fraction in the model parameters is numerically negligible (e.g., < 0.05). Multiple runs with different initializations are deployed to avoid local optima.
MotifHyades (Gibbs)

E-Step:

Although MotifHyades (EM) is the exact form of expectation maximization, it is noted that the overall computational complexities are proportional to the total sequence lengths which can scale and summed up to millions. Therefore, a variant is derived with the partial help of Gibbs Sampling, named as MotifHyades (Gibbs), as follows.

\[
P(x^{(t)} = t | D; \Theta) = \frac{P(PSeq^{(i)}, ESeq^{(i)} | x^{(t)} = t; \Theta) P(x^{(t)} = t; \Theta)}{\sum_{i=1}^{\infty} P(PSeq^{(i)}, ESeq^{(i)} | x^{(t)} = t; \Theta) P(x^{(t)} = t; \Theta)} \quad \forall t, i
\]

Sample \( z_{ip}^{(t)} \sim P(z_{ip}^{(t)} | x^{(t)} = t; D; \Theta) \) \quad \forall t, i

Sample \( z_{ik}^{(t)} \sim P(z_{ik}^{(t)} | x^{(t)} = t; D; \Theta) \) \quad \forall t, k

where samples are drawn for \( z_{ip}^{(t)} \) and \( z_{ik}^{(t)} \) using Gibbs Sampling. Therefore, \( F_{p}^{(i)}(i, j, c) \) and \( I_{p}^{(i)}(j, c, a) \) have been combined to a single function \( B_{p}^{(i)}(z_{ip}^{(t)}, j, c) \); \( F_{l}^{(i)}(i, j, c) \) and \( I_{l}^{(i)}(j, c, a) \) have been combined to a single function \( B_{l}^{(i)}(z_{il}^{(t)}, j, c) \). It reduces the computational complexities to be independent of sequence lengths \( L_{p}^{(i)} \) and \( L_{l}^{(i)} \).

All the model parameters \( \Theta = (M, P) \) are randomly initialized at the beginning. E-step and M-step are then alternated and repeated until the change fraction in the model parameters is numerically negligible (e.g., < 0.05). Multiple runs with different initializations are deployed to avoid local optima.
Proposed Approach (MotifHyades)
Simulation Results

• We have simulated artificial data for performance verification
  – Following the practice of motif discovery (Tompa), a series of simulation experiments with real DNA motif implantations are conducted for performance verification and comparison. In particular, the promoter and enhancer nucleotide background frequencies are estimated to generate $T$ random sequences (pmid22413003). The sequence lengths vary across multiple simulation scenarios. For each simulation experiment, 2K real human DNA motifs are downloaded either from the CIS-BP database (v1.01) (pmid25215497) or the JASPAR database (i.e. JASPAR_CORE_2016_vertebrates). The real motifs are randomly implanted into each simulated sequence pair as $K$ random DNA motif pairs. Given those $K$ known DNA motif pairs, we adopt the performance metrics MPD (Motif Pair Distance) and MPFR (Motif Pair Found Ratio) to evaluate the de novo motif discovery performance of different algorithms under diverse parameter settings of $K$ and $T$. 
Simulation Results – Motif Pair Distance (MPD)

MotifHyades (EM)
MotifHyades (Gibbs)
Wong et al. 2016

\[ MPD = \frac{1}{K} \sum_{i=1}^{K} \min(D(m^i_P, M^x_P) + D(m^i_E, M^x_E)) \]
Simulation Results – Motif Pair Found Ratio (MPFR)

MotifHyades (EM)
MotifHyades (Gibbs)
Wong et al. 2016

\[
MPFR = \frac{1}{K} \sum_{i=1}^{K} I[D(m^i_p, M^p_E) < 0.5 \land D(m^i_E, M^E_E) < 0.5]
\]
Simulation Results – Running Time (sec)

MotifHyades (EM)
MotifHyades (Gibbs)
Wong et al. 2016
Applications

- MotifHyades(EM) is run on the human promoter-enhancer-pair dataset (pmid27064255) of each human chromosome in six cell lines (i.e. K562, GM12878, HeLa-S3, HUVEC, IMR90, and NHEK) to discover 18,879 DNA motif pairs. The dataset is based on TargetFinder (pmid27064255). The count statistics of the discovered DNA motif pairs are visualized here. Examples are also depicted here. All 18,879 DNA motif pairs discovered can be downloaded openly.
Applications

- Those 18,879 motif pairs are characterized across a broad spectrum of genomic features on long-range promoter-enhancer pairs.

**Fig. 4.** Box plots on the DNase hypersensitivity peak fraction of the DNA motifs found on promoters and enhancers with motif pairing multiplicity. The solid blue lines are drawn using linear smooth function with 99% confidence intervals in R language (i.e. geom_smooth(method = "lm", level=0.99)). The sub-figures are sorted by cell type.
Applications

- Those 18,879 motif pairs are characterized across a broad spectrum of genomic features on long-range promoter-enhancer pairs.
Applications

• Those 18,879 motif pairs are characterized across a broad spectrum of genomic features on long-range promoter-enhancer pairs.
DNA Motifs on Chromatin Interactions [2]

- We have developed MotifHydes modeling for de novo motif pair discovery on paired sequences of promoters and enhancers on long-range chromatin interactions. Two model training algorithms have been developed for efficient model building with linear computational complexity.

- MotifHyades has also been adopted to discover 18,879 DNA motif pairs on the recently available promoter-enhancer regions in five other human cell types. Those 18,879 DNA motif pairs are characterized in extensive genomic contexts. The motif pairs and its genomic characterization can shed light on the complex long-range gene transcription process. It can be cross-validated with other sequencing studies, enabling thousands of further focused studies in the future.

Pattern Recognition on CRISPR-Cas9 Off-targets [3]

Pattern Recognition on CRISPR-Cas9 Off-targets [3]

CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes

Authors Authors and affiliations
Puping Liang, Yanwen Xu, Xio Zhang, Chenhui Ding, Rui Huang, Zhen Zhang, Jie Li, Xiaowei Xie, Yuxi Chen, Yujing Li, Ying Sun, Yaofu Bai, Zhou Songyang, Wenzhi Ma, Canqian Zhou, Junju Huang

Open Access Research article
First Online: 18 April 2015

ABSTRACT

Genome editing tools such as the clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (Cas) have been widely used to modify genes in model systems including animal zygotes and human cells, and hold tremendous promise for both basic research and clinical applications. To date, a serious knowledge gap remains in our
Pattern Recognition on CRISPR-Cas9 Off-targets [3]


One lab finds that a similar method produces large genetic deletions when performed in mouse embryos.

Two groups of researchers are questioning the results of an endeavor reported last year in which scientists edited the genomes of human embryos using CRISPR-Cas9. One team describes in Nature today (August 8) that when a similar method is applied to mouse embryos, their genomes end up with long deletions. The other critique lays out possible alternative explanations for the findings of the original study.

Pattern Recognition on CRISPR-Cas9 Off-targets [3]

CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences

Yanni Lin¹, Thomas J. Cradick¹, Matthew T. Brown¹, Harshavardhan Deshmukh¹, Piyush Ranjan², Neha Sarode², Brian M. Wile¹, Paula M. Vertino³, Frank J. Stewart² and Gang Bao¹,*

¹Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA, ²School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and ³Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA 30322, USA
CRISPR-Cas9 Off-targets

Motivation

Although specific fragments of DNA are aimed, sgRNA can sometimes influence other regions and incur off-target mutations (Chen et al. 2017). CRISPR-Cas9 can tolerate mismatches in sgRNA-DNA at different positions in a sequence-dependent manner, sensitive to the number, position, and distribution of mismatch (Hsu et al. 2013 and Zhang et al. 2015). Since sgRNA is able to endure some mismatches across several nucleotide positions, many off-target sites could be found on the target genome (Kim et al. 2015).

Off-target mutations can lead to genomic instability and disturb the normal gene function, which is still a major problem when applying CRISPR-Cas9 gene editing to clinical applications. Consequently, we still need accurate off-target prediction methods for complementary purposes.
Motivation

Most of the existing off-target prediction methods just calculate scores based on the positions of the mismatches to the guide sequence. The score of each base pair in sgRNA-DNA is derived using classic bioinformatics statistics on the mismatch effects based on previous gene editing experiments. For example, CFD (Cutting Frequently Determination) score is derived by infecting a large number of sgRNAs with single-bp replacement, deletion or insertion corresponding to the validated sgRNAs in MOLM13 cells; it calculates the percentage activity rates of different mutation sites based on LFC (Log Fold Change) value (Doench et al. 2016).

In light of the above, their performance are vulnerable to experimental variation. Most importantly, the existing methods can not take advantage of the growing CRISPR-Cas9 data for constant self-learning. In addition, most of the existing methods do not consider the potential relationships between mismatched and matched sites, which may be relevant to the off-target activity in CRISPR-Cas9 gene editing (Xu et al. 2017).
Proposed Approach (Deep Neural Network)

Fig. 1. An example on how to encode a sgRNA-DNA sequence pair. The table with thick borders in the middle of the figure shows the final matrix code of a sgRNA-DNA sequence pair, which can be used as the input for convolutional neural network modeling.

Fig. 2. The architecture of standard deep convolutional neural network (CNN_std) for off-target prediction. The input of this deep neural network is the encoded sgRNA-DNA sequence with length 23, which are converted into hidden features by one convolutional layer. The convolutional layer consists of 40 filters including 10 for each of the sizes $4 \times 1$, $4 \times 2$, $4 \times 3$ and $4 \times 5$. The BN layer is used to normalize the output of the convolutional layer in order to make the training process faster and avoid over-fitting. The global max-pooling layer applies a filter of window size 5 to the previous layers. The output of max-pooling layer are joined together into one vector by flattening. Each neurons in the flatten layer is fully connected to the first dense layer. Two dense layer consists of 100 and 23 neurons respectively. The second dense layer with a drop-out layer is fully connected to two output neurons to predict the probability of binary class. The neurons in output layer and dense layers use softmax function as the activation function, while all the neurons in other layers use ReLU as the activation function.
Performance Comparison on CRISPOR data

Fig. 3. ROC curves of two deep learning models and five current prediction methods under stratified 5-fold cross-validation on CRISPOR dataset.

Fig. 4. ROC curves of two deep learning models (i.e., FNN_3layer and DNN_std) and three traditional machine learning models including logistic regression (LR), random forest (RF) and gradient boosting trees (GBT). The ROC curve and AUC value of CFD score were regarded as the state-of-the-arts benchmark.
Performance Comparison on GUIDE-seq data

![ROC curves of deep learning models, CFD score and three traditional machine learning models on GUIDE-seq dataset.](image_url)
Fig. 6. 15 off-targets with the highest score predicted by final convolutional neural network and CFD score respectively on GUIDE-seq dataset. The sgRNA-DNA sequence marked with star is the true off-target.
We presented that deep convolutional neural networks are able to accurately predict the off-targets of CRISPR-Cas9 gene editing. To our knowledge, this is the first time that deep neural networks are designed and implemented for off-target predictions.

Our final convolutional neural network obtained the best performance on both CRISPOR dataset and GUIDE-seq dataset, outperforming the current state-of-art off-target prediction methods and three traditional machine learning algorithms including logistic regression, random forest, and gradient boosting trees. We discussed and attributed its performance successes to the neural network layer designs which are general enough to self-learn and capture sequence features by itself.

We believe that such intelligent approaches can contribute to CRISPR-Cas9 off-target predictions or other similar problems in a rigorous manner.

Pattern Recognition on CRISPR-Cas9 Off-targets [3]
Conclusion and Q&A

• Background

• Pattern Recognition on DNA Motifs:
  – DNA Motifs on Protein Binding Microarray [1]
  – DNA Motifs on Chromatin Interactions [2]

• Pattern Recognition on CRISPR-Cas9 Off-targets [3]

References


On-Going Research on Predicting DNA Motifs from DNA-Binding Protein Sequences
On-Going Research on Heterodimeric DNA Motif Synthesis and Validation

**Figure 1.** Overview of Proposed Approach. The approach is divided into 3 phases with distinct objectives. It is noted that the monomeric DNA motif of EOMES has been reverse-complemented for clarity.