EVALUATING MULTIPLE SEQUENCE ALIGNMENTS USING ENCODE DATA

A Thesis in
Integrative Biosciences
by
Qingyu Wang

© 2013 Qingyu Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2013
The thesis of Qingyu Wang was reviewed and approved* by the following:

Webb C. Miller  
Professor of Biology and Computer Science and Engineering  
Thesis Advisor

Ross Hardison  
T. Ming Chu Professor of Biochemistry and Molecular Biology

Naomi S. Altman  
Professor of Statistics

Peter J. Hudson  
Willaman Professor of Biology  
Chair of the Integrative Biosciences Program

*Signatures are on file in the Graduate School
ABSTRACT

As a key aspect of decoding genomic sequences, comparative genomic research has become increasingly prominent during the last decade. A crucial prerequisite for comparative genomics is multiple genomic sequence alignments, especially whole-genome multiple sequence alignments (MSAs). Various downstream analyses rely implicitly on whole-genome MSAs. Unfortunately, whole-genome MSAs still suffer from inadequate reliability and the research on assessing their quality has not been fully addressed yet. Regardless of whether we seek better alignment methods or want to select the most reliable available MSAs, a practical evaluation method for MSAs is imperative. Therefore, we propose a new method, MSAME (MSA Motif-based Evaluation), to quantify the reliability of MSAs based on the ChIP-seq data produced by the ENCODE project. Our method is one of the first MSA evaluation methods based on experimental data. Instead of using simulations relying on evolutionary models, we define a biological criterion inferred from different scenarios of motif shifting in ChIP-seq peak regions, and evaluate MSAs from such a functional perspective. Our method efficiently identifies two types of MSA errors, and it is robust to noises introduced by parameter changes. By applying our method to evaluate high coverage 11-way eutherian MultiZ and EPO alignments, we identify 7.9% and 6.7% of alignments in specific ChIP-seq bound regions as unreliable MSAs respectively. This allows us to identify putative evolutionary motif shifting as a by-product. Finally, we conduct further analysis for putative evolutionary motif shifting events that we detect for binding sites of the transcription factors GATA1, CTCF, and NRSF.
# TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. v

LIST OF TABLES ..................................................................................................................... vii

ACKNOWLEDGEMENTS ......................................................................................................... viii

Chapter 1  Introduction ........................................................................................................... 1

1.1 Background and significance ....................................................................................... 1
1.2 A short review of MSA evaluation methods ............................................................. 2
1.3 Motivation ................................................................................................................... 4

Chapter 2  Evaluating MSA Using ENCODE Data ............................................................. 7

2.1 Basic definition ........................................................................................................... 7
  2.1.1 Multiple sequence alignments ........................................................................... 7
  2.1.2 Transcription factor binding sites and ChIP-seq peaks ................................... 8
  2.1.3 Motifs and the position weight matrix .............................................................. 9
2.2 Overview ................................................................................................................... 11
2.3 Methods and materials ............................................................................................. 15
  2.3.1 Pseudo-code and a flowchart for our method ............................................... 15
  2.3.2 Detecting motif shifting events ...................................................................... 18
  2.3.3 Generating the target interval and the query interval ...................................... 18
  2.3.4 Local alignment statistics .............................................................................. 20
  2.3.5 Training scoring matrices and estimating Karlin-Altschul parameters ........... 20
  2.3.6 Significance test for non-homologous sequences ........................................... 21
  2.3.7 The test to classify a motif as either a putative motif shifting or product of local misalignment ............................................................... 22
  2.3.8 Unreliability score for MSAs surrounding a binding site .............................. 23
  2.3.9 Application to genome-wide alignments ...................................................... 24
2.4 Results and Evaluation .............................................................................................. 26
  2.4.1 Effects of changing the motif detection threshold .......................................... 27
  2.4.2 Evaluating MSAME using the datasets in Schmidt’s paper .......................... 28
  2.4.3 MSAME evaluation for MultiZ and EPO MSAs ............................................ 31
  2.4.4 Putative evolutionary motif shifting analysis ............................................... 36

Chapter 3  Discussion and Future work ........................................................................... 39

3.1 Discussion ................................................................................................................... 39
3.2 Future work ............................................................................................................... 41

REFERENCES CITED ......................................................................................................... 42
LIST OF FIGURES

Figure 2-1: Motifs, indicated by shaded blocks, in homologous sequences should be aligned together at the same position in the MSA block. This figure shows scenarios where motifs in non-human species are aligned in different positions in MSAs; in this paper we call each of these a “motif shifting event”. Three possible reasons are responsible for the occurrence of motif shifting events: (A). Local misalignment. The left alignment block is the original alignment, in which the motifs positions of mouse and rat are shifted from motif positions in other species. The right alignment block is the alignment after shifting the mouse and rat sequences to make their motif aligned with other species and then adjust gaps. It is obvious that the after-shifting alignment on the right has better alignment quality. Thus we could infer that this motif shifting is caused by local misalignment. (B). Non-homologous sequences. In this alignment block, cow aligns weakly to other species. It’s highly likely that the sequence of cow was aligned to this region incorrectly, thereby introducing an extra false motif site. (C). Putative evolutionary motif shifting. The motifs are supposed not to be aligned with other species. The overall alignment is good and the alignment is not improved by adjusting motif positions. .......................................................... 14

Figure 2-2: An example of aligned motifs and neighboring motifs. Given two human motifs, 1 and 2, in a peak region, blocks in grey colors are the aligned motifs, blocks in red and green colors indicate neighboring motifs for the human motif 1, and blocks in green and blue colors represent neighboring motifs for the human motif 2. .... 16

Figure 2-3: Flowchart of our method and result analysis. The whole process is composed of two steps: 1) train scoring matrices and estimate Karlin-Altschul parameters from MSAs without motif shifting events, and 2) test each motif shifting event related to human binding sites to identify unreliable MSAs. .............................................................. 17

Figure 2-4: Generating the consensus sequence as the target interval from the set of target sequences, and using of sequences with shifted motifs as the query intervals. Motifs in the shaded blocks will be used as the anchor in the realignment process. ...................... 19

Figure 2-5: A demonstration for generating the consensus sequence in order to train the score matrix. Gray blocks represent motifs detected on different sequences. .................. 21

Figure 2-6: (A) Background alignment: original alignment with motif shifted from positions of motifs in other species (B) Shifting alignment. Alignment after shifting the motif back to make it aligned to motifs in other species................................. 23

Figure 2-7: Result of running MSAME on (A) MultiZ and (B) EPO MSAs by changing the threshold cutoff for GATA1 detection using a PWM. Both MSAs show the same pattern of response to the change of cutoff. ................................................................. 28
Figure 2-8: An example of deeply conserved binding sites in five species (grey blocks) surrounded by lineage-specific (green blocks in rodents) and species-specific (red blocks) motifs presented in Schmid’s paper. 

Figure 2-9: A weak alignment was identified as a highly conserved region in Schmidt’s paper, but classified as non-homologous sequences by MSAME. Lower case letters identify bases that differ from human.

Figure 2-10: (A) The number of species covered in different MSAs in all peak regions of three transcription factors. (B) Conservation scores for three transcription factors in MSAs in their ChIP-seq peak regions and corresponding random genomic intervals.

Figure 2-11: Evaluation results on MSAs pre-computed by MultiZ and EPO in GATA1, CTCF, and NRSF ChIP-seq peak regions. (A) Percentage of binding sites with unreliable MSAs and percentage of binding sites with at least one motif shifting event in both MultiZ and EPO. UCSC has a high percentage of binding sites identified with unreliable MSAs partially due to a high percentage of binding site with motif shifting. (B) Distribution of unreliability scores for unreliable MSAs in MultiZ and EPO. (C) Percentage of binding sites with unreliable MSAs in GATA1 ChIP-seq peak regions in four different region categories: TSS upstream, TTS downstream, intronic and intergenic regions. (D) Percentage of motif shifting events classified as different categories in MultiZ and EPO: non-homologous sequences, local misalignment, and putative evolutionary motif shifting. MultiZ has a slightly lower rate of non-homologous sequences. (E) Distribution of motif shifting identified as non-homologous in different species in MSAs from MultiZ and EPO. Except dramatically high percentage in dog, MultiZ has a lower or similar rate in other species than EPO.

Figure 2-12: Different patterns of the percentage of binding sites with putative evolutionary motif shifting in GATA1, CTCF, and NRSF ChIP-seq peak regions.

Figure 2-13: Putative evolutionary motif shifting in different species. (A) Distribution of putative evolutionary motif shifting occurring in GATA1 binding sites. (B) Distribution of distances that a motif shifted in different species in MultiZ MSAs, away from the position corresponding to the human binding site.
LIST OF TABLES

Table 2-1: One of the position-specific weight matrix for the transcription factor GATA1 downloaded from JASPAR database ......................................................... 10

Table 2-2: A position-specific probability matrix for a motif with length n = 4. ................. 11

Table 2-3: A position weight matrix for a motif with length n = 4 using background probabilities A=0.3, C=0.2, G=0.2, T=0.3 ................................................................. 11

Table 2-4: The PFM we used for mapping the motif of GATA1 ........................................ 25

Table 2-5: The PFM we used for mapping the motif of NRSF ........................................ 25

Table 2-6: The PFM we used for mapping the motif of CTCF ........................................ 25
ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Webb Miller, my thesis and academic advisor, for his expertise and encouragement throughout my entire educational process. He gave me the opportunities to choose my own research interests and to pursue exciting academic research projects. Without his generosity and his patient editorial comments, suggestions, and revisions, this thesis would not been possible.

I also would like to thank my committee members, Dr. Ross Hardison and Dr. Naomi Altman. Dr. Hardison gave me useful suggestions and comments for my thesis research and provided me with the chance to present my work in his lab meetings. Dr. Altman gave me valuable experience in solving statistical problems and offered me helpful suggestions on my graduate study and future career path. I also would like to express gratitude to Dr. Cooduvalli Shashikant, the chair of the IBIOS-BG graduate program, who encouraged and supported me throughout the development of my thesis.

I must acknowledge Kuan-Bei Chen, whose intelligent conversations and insightful feedback assisted me greatly. I would also like to thank my research mates Oscar Bedoya Reina, Giltae Song, and Aakrosh Ratan.

I am forever indebted to my parents and my husband, Hai Zhu, for the all of the support they have given me throughout the years. Without their love and understanding, I would not be where I am today.
Chapter 1

Introduction

1.1 Background and significance

With the increasing number of complete genome sequences that have become available for different species during the past decade, comparative genomic analysis plays more and more significant roles for annotating functional elements. In all comparative genomic methods, the most fundamental and crucial component is a multiple genomic sequences alignment, especially a whole-genome multiple sequence alignment (MSA). A variety of downstream analyses, including (but not limited to) annotating protein-coding genes (Gross et al. 2006), RNA genes (Washietl et al. 2005), and regulatory sites in non-coding regions (Dermitzakis et al. 2002, Daily et al. 2011), predicting which genomic differences are functional, and constructing phylogenetic trees (Delcher et al. 2005), are based on whole-genome MSAs.

Unfortunately, despite being widely used in various research efforts, such MSAs as MultiZ (Blanchette et al. 2004) and PECAN (Paten et al. 2008) still suffer from both inadequate accuracy and reliability, as well as lacking a widely accepted assessment of quality. Part of the reason for this is the inadequate evolutionary models on which the alignments are based. Yet, the major reasons actually are the NP-hard nature of computing exact optimal MSAs (Wang et al. 1994, Elias 2006) and the large search space due to massive genomic rearrangement (Praksh et al. 2007). Given such limitations, almost all current whole-genome MSA programs use heuristic (mostly tree guided) methods to compute sub-optimal alignments. However, since enumerating all sub-optimal alignments is impractical (Naor et al. 1994), current programs only report a single, somewhat arbitrary sub-optimal alignment, which introduces uncertainty in the
alignments. Additionally, false homology inference caused by large scale genomic rearrangements tends to introduce unrelated sequence into the alignment. As a result, a considerable disagreement has been observed in alignments generated by different MSA programs. Moreover, due to lack of a quantitative way to measure the reliability of alignments, further downstream analyses are jeopardized. From this perspective, the quality of MSAs to a great extent determines the accuracy of comparative genomic studies. Therefore, a practical measurement for the quality of whole-genome MSAs is essential, an issue that has been addressed recently by scholars (Parkash et al. 2005, Margules et al. 2007, Chen et al. 2010).

1.2 A short review of MSA evaluation methods

The development of evaluation methods for MSAs has lagged behind design efforts for MSA algorithms. Unlike protein sequence alignments, which can be evaluated using experimentally determined protein structures and highly curated benchmark databases (Parkash et al. 2005), there are no stringent “true” standards for genomic sequence alignments (Margules et al. 2007). Consequentially, evaluating whole-genome MSAs becomes a very challenging task. Most current assessment methods are based either on inferences from a certain evolutionary model or putative “true” biological alignments. Both TBA and PECAN used simulation to generate artificial biological sequences based on models for evolution of neutral genomic regions, and compared their alignment result to those simulated “true” biological correspondences. Such simulation methods assume that simulated sequences are similar to real biological sequences. However, due to inadequate evolutionary models and the complexity of evolutionary events, such as conversion and duplication, generating biological-like sequences is very challenging. MLAGAN (Brudno et al. 2003) and MAVID (Bray et al. 2004) were assessed by running the aligner on existing orthologous sequences to see if the protein-coding exons in those sequences
are aligned correctly. Although such biological-sequences-based evaluation methods provide a
direct biological criterion for the quality of MSAs, the performance evaluation is limited by the
possible inaccuracy of orthologous prediction and the high similarity of test aligned regions. To
avoid these drawbacks, Margules’ (Margules et al. 2007) introduced the concept of MSA
specificity by calculating the rate at which human Alu sequences are aligned to genomes that lack
Alus, as well as measuring levels of periodicity in coding exons. However, only alignments in
certain regions could be evaluated by those methods.

More recently, several independent tools were developed for evaluating MSAs.
StatSigMA-w (Prakash et al. 2007) extended the Karlin-Altschul statistics (Karlin et al. 1990)
from pair-wise alignment to multiple local alignment statistics to identify regions being
contaminated with unrelated species, based on a certain phylogenetic tree. An advantage of this
method is to quantify the accuracy of whole-genome MSAs at the nucleotide level, and it can be
applied to any genomic regions. However, the performance of this method still highly depends on
phylogeny and context correctness, which might be difficult to achieve. GUIDANCE (Penn et al.
2010) provides a web server for the assessment of MSAs; two algorithms, HoT (Landan et al.
2007) and GUIDANCE (Landan et al. 2008), were implemented and integrated in this
application. The HoT algorithm is rooted on the assumption that the result of alignment is
sensitive to the direction of input sequences. Therefore, the measurement of quality of MSAs
could be inferred by comparing the heads (alignments with original direction) and tails
(alignments with reversed direction) alignment. The method is intuitive and easy to implement.
However, later research (Wise 2010) suggested that HoT quality scores are not sufficient to
evaluate MSAs due to limitations of sampling all possible sub-optimal alignments. The HoT
algorithm was then extended to apply to a set of co-optimal MSAs constructed by a progressive
method (Landan et al. 2008). GUIDANCE was based on the assumption that the inaccuracy of
MSAs is mainly caused by uncertainties of the guide tree. A set of perturbed MSAs was
generated by using the progressive method based on bootstrap guide trees. By comparing those perturbed MSAs with the original MSAs, a confidence score was assigned to each base. The simulation and benchmark test showed GUIDANCE has better performance in capturing alignment errors than does HoT. However, since during the process of constructing perturbed guide trees and alignments the putative misaligned sequences were taken into account as well, the reliability of the confidence scores based on such agreement of perturbed alignment and original alignments is imperfect. The most recently released tool was PSAR (Kim et al. 2011), which is based on an assumption that the reliability is reflected in the sub-optimal alignments. Therefore, PSAR generates a large set of sub-optimal alignments by a probabilistic sampling method. The agreement between the input MSAs and its sub-optimal set gives the reliability scores. Simulation and benchmark results showed that PSAR performs better than GUIDANCE. Also, as a bonus, the set of sub-optimal alignments can be used for downstream analyses. On the other hand, since PSAR generates a large set of sub-optimal alignments directly without using any heuristic assumptions, the computational time is relative long, and it may not be feasible to evaluate MSAs with a large number of species.

1.3 Motivation

Based on the requirement of downstream analyses, whole-genome MSAs generally served two purposes, related to either evolution or biological function. In some cases, functionally correct alignments are not necessary evolutionarily related. Therefore, evaluation methods should be addressed from both these two perspectives. Even so, most of the existing assessment criteria, including the methods we discussed above, are focused on evaluating evolutionary correctness of the MSAs, based on different evolutionary assumption. However, almost all MSAs were constructed based on the phylogenetic guide tree, so evaluation based on phylogenetic trees or an
evolutionary model implies some circularity. Therefore, in this paper, we propose a new MSAs evaluation method to measure accuracy from the functional-based perspective, by using ENCODE data.

Sequence alignment is not only a pure computational challenge, but far more a biological problem. To this end, the ideal evaluation method and criteria applied should be based on real biological benchmark data, such as BAibase (Thompson et al. 2005) for protein MSAs. Although, as we discussed above, such a benchmark database is not practical for genomic MSAs, a variety of functional experimental data available from the ENCODE project provide us with an alternative biological “benchmark”, which can be applied for assessing MSAs. Aiming for identifying all functional elements in the human genome, the ENCODE project has already associated 80% of the components of the human genome with at least one biochemical function (Dunham et al. 2012). As one of the significant parts, ChIP-seq data from the ENCODE project provide substantial information about the sequence features around the transcription factor (TF) binding regions. For instance, binding regions for 87 sequence-specific transcription factors are assayed by ChIP-seq in the ENCODE project already, and more than 80% of them are associated with strong DNA-binding motifs (Wang et al. 2012). Additionally, since the conservation of transcription factors, such as GATA1 and CTCF, among vertebrate species has long been studied, it is feasible to consider the combination of both ChIP-seq bound regions and motif sequence features as a biological criterion for MSAs evaluation.

Herein, we presented a program, MSAME (MSA Motif-based Evaluation), to quantify the reliability of MSAs based on ChIP-seq data produced by the ENCODE project. MSAME detected potential misalignment in MSAs around transcription factor binding sites and categorized them based on two types of MSA errors. Our method is based on the functional perspective of alignment evaluation. We hope that, from this method, we could define a biological criterion for the measurement MSA accuracy by applying available experimental data,
especially ChIP-seq data, produced by the ENCODE project. We apply our method to high coverage 11-way eutherian MultiZ alignments and EPO alignments from UCSC and Ensembl respectively and compared the quality of these two alignments in specific ChIP-seq bound regions. As ChIP-seq data is available from a very limited set of species (mostly just human) in the ENCODE project, in this paper we only used human ChIP-seq data as strong evidence for the occurrence of binding sites motifs. Therefore, only peak regions with motif sequences detected in human will be taken into account for the evaluation. Below, “binding sites” refers to binding site motifs identified in human sequence only. For computationally predicted motif sequences mapped in other species in MSAs, we will simply use the term “motif” to distinguish them from experimentally validated motifs in human. Our result showed that, overall, EPO alignments have a lower percentage of unreliable MSAs (6.7%) than MultiZ (7.9%). However, MultiZ has a lower rate of non-homologous sequences in most of species than does EPO. Since this evaluation method is based on motif information, by filtering the potential misalignment we could apply this method to identify conserved motifs not on the preserved position, i.e., putative evolutionary motif shifting (meaning that motifs are still functionally conserved, but their positions are shifted in different species during evolution). We therefore conducted further analysis for such putative evolutionary motif shifting occurred around transcription factors GATA1, CTCF, and NRSF binding sites.
Chapter 2

Evaluating MSA Using ENCODE Data

2.1 Basic definition

2.1.1 Multiple sequence alignments

A multiple sequence alignment (MSA) is an alignment including more than two biological sequences from different species. More precisely, a MSA is a transformation of $k > 2$ biological sequences $\{S_1, S_2, \ldots, S_k\}$ into a set of mapped strings $\{S'_1, S'_2, \ldots, S'_k\}$ that may contain dashes (indicating insertions or deletions), where $S'_1, S'_2, \ldots, \text{and } S'_k$ are in the same length, the removal of dashes from $S'_i$ leaves $S_i$ for each $1 \leq i \leq k$, and at each position the symbols of $S'_1, S'_2, \ldots, \text{and } S'_k$ are not all dashes. If such alignment has been done to each position of the sequences, we call it the global alignment; otherwise, if only the most similar parts of the sequences are aligned, we call it the local alignment. Generally, biological sequences being aligned can be either genomic DNA sequences or protein sequences. In this article, we focus on alignments of genomic DNA sequences, especially whole genome-wide DNA sequences, from multiple species.

Although we are discussing whole genome alignments here, due to the presence of complex genomic structures it is generally impractical to globally align the entire genome sequences in multiple species. Moreover, large scale genomic rearrangements, such as inversion and conversion, create non-collinearly among the corresponding functional regions in different genomes. Therefore, a set of local alignments of homologous regions, also called a set of alignment blocks, is what we expect to acquire from aligning genome-wide sequences. Most
current genome-wide MSA programs, such as TBA, Pecan, MLAGAN, and MAVID, are designed based on such strategy, beginning with inferring a set of homologous sequences and followed by applying the specific multiple sequence aligner, either local, global or both, to generate final MSAs. A more complicated case than genomic rearrangement is segmental duplication, which requires the aligner to be capable of aligning the given position in one genome to multiple positions in other genomes. With the exception that MAVID requires strict one-to-one orthology, the aligners mentioned above to some extent allow a given genomic position to be aligned more than once. However, we feel that the quality of currently available MSAs in such duplicated regions is not good enough. Both rearrangements and duplications bring a potential risk of introducing unrelated sequences into MSAs.

To produce theoretically optimal MSAs, a direct method is multi-dimension dynamic programming. However, since the run-time complexity of multi-dimension dynamic programming is \(O(2^n n^k)\) (k sequences of length n) (Durbin et al, 1998), computational time of this algorithm will grow exponentially with the number of sequences, which is impossible to be applied to whole-genome data with many sequences. Most current existing MSA programs use heuristic methods, generally based on phylogenetically guided progressive methods, to build MSAs from pair-wise alignments (Brudno et al. 2003, Blanchette et al. 2004). The alignments produced from such methods are therefore suboptimal alignments and the uncertainty is introduced to MSAs during the process of arbitrarily choosing from among the suboptimal alignments.

### 2.1.2 Transcription factor binding sites and ChIP-seq peaks

A transcription factor is a protein that binds to a specific DNA sequence in regulatory regions of genes to regulate the gene’s expression. A genomic interval that under certain
conditions is bound by a transcription factor is called a transcription factor binding site (TFBS). Generally, a TFBS is from 5bp to 31bp long, with a typical length around 10bp (Stewart et al, 2012). One efficient experimental method to identify TFBSs is ChIP-sequencing (ChIP-seq), which combines chromatin immunoprecipitation (ChIP) experiments with massively parallel DNA sequencing (Robertson et al. 2007). In the ChIP-seq assay, genomic DNA fragments cross-linked with transcription factors are immuno-precipitated by an immune reagent specific to the factor, followed by sequencing the transcription factor bound genomic fragments and aligning those reads back to the reference genome sequence. The sites with a local concentration of high read counts identified by peak-calling programs are called ChIP-seq peaks (Johnson et al. 2007), which are regions containing putative TFBSs. Since the average size of ChIP-seq peak regions is around few hundred bp, one or more putative TFBSs can be found in a ChIP-seq peak region. Research show that TFBSs tend to aggregate around ChIP-seq peak summits, and TFBSs in ChIP-seq peak regions are highly conserved evolutionarily (Wang et al, 2012).

2.1.3 Motifs and the position weight matrix

The common sequence pattern shared by specific DNA sequences in binding sites for a transcription factor is defined as the factor’s motif, also called a transcription-factor binding motif. By detecting motifs in genomic DNA sequences using a computational method, putative TFBSs can be predicted for a specific transcription factor. One of the more efficient methods to model and map motif patterns uses a position weight matrix (PWM), also called a position-specific scoring matrix (PSSM) (Wasserman et al. 2004). For a motif with length L, a PWM is an L by 4 matrix giving a weighted match to each position of the motif at each base. PWMs for motifs corresponding to many transcription factors can be converted from position frequency matrices (PFMs) downloaded from PWM databases, such as JASPAR (Sandelin et al, 2004) and
TRANSFAC (Matys et al, 2003) produced by BIOBASE. The PFM records the total number of observations of each nucleotide for each position of the motif. There could be multiple different motif sequences with corresponding PFMs and PWMs for one transcription factor. Table 2-1 shows one of PFMs from JASPAR for the motif of the transcription factor GATA1.

Table 2-1: One of the position-specific weight matrix for the transcription factor GATA1 downloaded from JASPAR database

<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>1423 708 2782 0 4000 27</td>
<td>3887 3550 799</td>
<td>1432 1487</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>560 1633 31 0 0 29 0 4</td>
<td>681</td>
<td>897</td>
<td>829</td>
</tr>
<tr>
<td>G</td>
<td>1242 1235 10 4000 0 109 6</td>
<td>383</td>
<td>2296 1360</td>
<td>1099</td>
</tr>
<tr>
<td>T</td>
<td>775 424 1177 0 0 3835 107</td>
<td>63 224 311</td>
<td>585</td>
<td></td>
</tr>
</tbody>
</table>

Given a PFM, we can convert it to a position-specific probability matrix (PSPM), where the value of each cell, $p_{ij}$, is the probability of observing nucleotide $i$ at position $j$ of the motif. A PWM, then, can be formed by calculating a log ratio for each cell of the PSPM based on the background probability of seeing each nucleotide in the background sequence. The equation to compute such log ratio is: $\log\left(\frac{p_{ij}}{b_i}\right)$, where $b_i$ is the background frequency of nucleotide $i$. Using a PWM, new sequences can be scored by summing the corresponding values of the observed base in each position, $S = \sum_{i=0}^{L-1} \log\left(\frac{p_{ij}}{b_i}\right)$.

For example, consider testing the sequence “AATC” with the toy background probabilities A=0.3, C=0.2, G=0.2, T=0.3 and the toy PSPM of length $n = 4$ shown in Table 2-2. The PWM is generated as in Table 2-3. The score of observing “A” in the first position based on this PWM is -0.08, “A” in the second position is 0, “T” in the third position is -1.08, and “C” in the fourth position is 0.63. The sum of scores, $S = -0.08 + 0 + (-1.08) + 0.63 = -0.53$, is the PWM score of the sequence “AATC” being an instance of the motif. To measure the quality of match based on the PWMs in different lengths, this score is normalized by the formula $\frac{S-S_{\min}}{S_{\max}-S_{\min}}$ and the normalized score is ranged between 0 and 1. Sequences likely to be instances of the motif can be determined based on a cutoff for such normalized scores.
2.2 Overview

Many MSA errors fall into one of two categories: uncertainty in local alignments due to arbitrarily choosing sub-optimal alignments, and unrelated sequences incorrectly aligned due to false homology inference. MSAME evaluates MSAs from these two perspectives based on the occurrences of transcription factor binding motif sequences in different scenarios observed in MSAs. Because they perform a biological function related to gene regulation, motif sequences are known to evolve at a relatively lower rate than non-functional regions. Therefore, conserved
motifs in homologous regions are expected to be aligned together, and various programs predict de novo motif candidates from MSAs based on such rules (Daily et al. 2011). Even so, cases in which the accuracy of such a prediction was effected by misalignment have been observed (Gordan et al 2010, Ray et al. 2008). Due to the two types of errors in MSAs, either the positions of motifs are shifted from the aligned position, as in Figure 2-1A, or unreliable motifs from unrelated sequences are introduced in the alignment as in Figure 2-1B. In both cases, motif shifting – cases in which motifs in different species are not aligned together in the homologous sequences – is observed in the MSA. Such motif shifting increases the difficulty of predicting motifs; on the other hand, it gives us a biological perspective to identify unreliable alignments. The occurrence of motif shifting in MSAs can be applied to infer their quality.

Nevertheless, although unreliable MSAs can cause motif shifting from aligned positions, the serious concern here is that it is not the only reason for the occurrence of such motif shifting. Owing to the potential mobility of transcription factor binding sites during genomic evolution, large numbers of functionally conserved motifs do not align between species (Odom et. al, 2007), as shown in Figure 2-1C. To this end, distinguishing whether a motif shifting is caused by a “true” evolutionary motif shifting event or by the MSA’s unreliability becomes a very challenging task, and few research efforts have addressed this issue.

MSAME infers statistical significance for solving this issue, thereby identifying the motif shifting caused by unreliable alignments, based on the motif-anchored realignment and subalignment significance tests, as described below. A strong assumption behind our method is that the TFBSs in ChIP-seq peaks are deeply conserved, not only in motif sequences but also in their flanking sequences. Studies in some transcription factors, such as GATA1 (Vyas et al. 1999) and CTCF (Schmid et al. 2012), supported such assumption. Based on this assumption, the null hypothesis used to motivate our method is that the flanking sequence of the shifted motif is unrelated to the flanking sequences of a set of homologous sequences (target sequences or
consensus sequence, as we will discussed below) anchored by such position. We addressed “anchor” here because our interest is in determining if the binding motif sequence at the specific position has been well aligned, but not any aligned sub-sequence without the motif sequence. Based on this hypothesis, we conducted one significant test for ruling out motif shifting caused by introducing non-homologous sequence, followed by the second test, to distinguish between putative evolutionary motif shifting and local misalignment. For each observed case of motif shifting between species, we computed these two kinds of p-values, classified each motif shifting event into three categories, and assigned them a score based on the test result as shown blow.

\[
Motif\ Shifting \Rightarrow \begin{cases} 
  Misalignment & s = 0.05 \\
  Local\ misalignment \quad Non-\ homologous\ sequences & s = 1 \\
  Putative\ evolutionary\ motif\ shifting & s = 0 
\end{cases}
\]
Figure 2-1: Motifs, indicated by shaded blocks, in homologous sequences should be aligned together at the same position in the MSA block. This figure shows scenarios where motifs in non-human species are aligned in different positions in MSAs; in this paper we call each of these a “motif shifting event”. Three possible reasons are responsible for the occurrence of motif shifting events: (A). Local misalignment. The left alignment block is the original alignment, in which the motifs positions of mouse and rat are shifted from motif positions in other species. The right alignment block is the alignment after shifting the mouse and rat sequences to make their motif aligned with other species and then adjust gaps. It is obvious that the after-shifting alignment on the right has better alignment quality. Thus we could infer that this motif shifting is caused by local misalignment. (B). Non-homologous sequences. In this alignment block, cow aligns weakly to other species. It’s highly likely that the sequence of cow was aligned to this region incorrectly, thereby introducing an extra false motif site. (C). Putative evolutionary motif shifting. The motifs are supposed not to be aligned with other species. The overall alignment is good and the alignment is not improved by adjusting motif positions.
2.3 Methods and materials

2.3.1 Pseudo-code and a flowchart for our method

Our method works in two steps: 1) train alignment scoring matrices and estimate parameters for inferring statistical significance of pairwise alignments, and 2) test each motif shifting event to identify unreliable MSAs using trained parameters from the first step. One assumption behind our method is that we consider all sequences with aligned motifs in MSAs to be putative homologous sequences. Based on this assumption, by detecting motif shifting events, we divided MSAs in all ChIP-seq peak regions into two groups: the training set and the testing set. MSAs with no motif shifting are used as the training set to acquire important parameters; MSAs with motif shifting are tested to detect unreliable MSAs. Our method takes MAF (Multiple Alignment Format) files, interval files that define ChIP-seq peak regions, and PWM files as input. We summarize the method using both pseudo-code and a flowchart (Figure 2-3). Each step will be further discussed in later sections.

Before giving the pseudo-code, we need two definitions. Given a human motif in a peak region, an *aligned motif* is a motif starting in precisely the same columns as the human motif (including the human motif itself). A motif in a species that does not have an aligned motif is a *neighboring motif* for the human motif if:

1. the motif aligns within the same human ChIP-seq peak region and
2. no other human motifs align between the non-human motif and the place where the original human motif aligns.

These definitions are illustrated in Figure 2-2.

Pseudo-code to detect and test motif shifting events:

1. For each motif in a human ChIP-seq peak region do
   1.1. For each neighboring motif for this human motif do
1.1.1. Form a “target interval” (at least 200bp) as the consensus for all rows of the MSA with an aligned motif
1.1.2. Form a “query interval” (same length as the target interval) in this non-human species
1.1.3. For each position within the query interval do
   1.1.3.1. Use the motif sequence in the target interval as the seed to align to the position and extend on both sides ignoring gaps, calculating $p(\text{position}) = p$-value of the gap-free local alignment with the maximal score
   1.1.3.2. If position = start position of the neighboring motif then $p(\text{background}) = p(\text{position})$
   1.1.4. $P \leftarrow \min\{p(\text{position}) : \text{positions in query sequence}\}$
   1.1.5. If $P \geq 0.05$ then Motif shifting is introduced by non-homologous sequence
   1.1.6. If $P < 0.05$ and $p(\text{background}) \geq 0.05$ then Motif shifting is caused by local misalignment
   1.1.7. If $P < 0.05$ and $p(\text{background}) < 0.05$ then Motif shifting is putative evolutionary motif shifting and MSA is reliable

Figure 2-2: An example of aligned motifs and neighboring motifs. Given two human motifs, 1 and 2, in a peak region, blocks in grey colors are the aligned motifs, blocks in red and green colors indicate neighboring motifs for the human motif 1, and blocks in green and blue colors represent neighboring motifs for the human motif 2.
Figure 2-3: Flowchart of our method and result analysis. The whole process is composed of two steps: 1) train scoring matrices and estimate Karlin-Altschul parameters from MSAs without motif shifting events, and 2) test each motif shifting event related to human binding sites to identify unreliable MSAs.
2.3.2 Detecting motif shifting events

We identified transcription factor binding sites and motifs by applying positional weight matrices (PWMs) to sequences within MSAs in human ChIP-seq peak regions. Since our data is highly human centric, only peak regions with at least one motif detected in human will be taken into consideration. Thereafter, motif shifting events, or neighboring motifs defined previously, will be detected as cases where a motif in a non-human species occurs in a different position compared to human in the MSA under consideration. As illustrated in Figure 2-4, we took the set of sequences with motifs occurring at the same position in binding sites, namely aligned motifs, as the target sequences, and sequences with motif shifting as the query sequences. Human sequence must be included in the target sequences set. We assumed that the target sequences were homologous sequences because motif sequences were already aligned together. The question we would like to answer is: might that the positions of motifs in the target sequence shift from those in consensus sequence is the result of MSA errors?

2.3.3 Generating the target interval and the query interval

Since MSA errors are introduced during the MSA construction process, the consensus sequence was generated from the target sequences to avoid and eliminate such bias in the evaluation method. A target interval and a corresponding query interval containing the neighboring motif being tested are generated based on the relative positions of aligned and neighboring motifs. The two are aligned and statistical significance is inferred, as demonstrated in Figure 2-4. To cover all the alignment information of the sequences including the aligned motifs and the tested neighboring motif, the leftmost start position of the aligned motifs and the tested neighboring motif was defined as the left boundary, and the rightmost end position of the aligned
motifs and the tested neighboring motif was defined as the right boundary. The target interval and the query interval were generated as the sequence between the right boundary and the left boundary plus upstream 100bp flanking sequence of the left boundary and the downstream 100bp flanking sequence of the right boundary. Thus, the length of the target sequence and the query sequence generated based on such consensus interval and query interval are at least 200bp, varying for different tested neighboring motifs. By this step, the MSA evaluation problem is transformed to a problem of comparing two sequences. Thus, we can make a statistically valid inference for such pairwise alignment between the target sequence and the query sequence. Meanwhile, the major content information at each position in all the target sequence sets is maintained. Additionally, as the uncertainty of MSAs is generally due to incorrect introduction of gaps (Durbin et al. 1998, Landan et al. 2009), we used gap-free segment scores for estimating local alignment significance.

Figure 2-4: Generating the consensus sequence as the target interval from the set of target sequences, and using of sequences with shifted motifs as the query intervals. Motifs in the shaded blocks will be used as the anchor in the realignment process.
2.3.4 Local alignment statistics

The statistical significance of the subalignment of two random sequences without gaps can be calculated by Karlin-Altschul statistics, as follows (Altschul, 1997).

\[ Prob(S \geq x) = 1 - \exp(-Kmne^{-\lambda x}) \] (1)

We extended this statistic to evaluate MSAs around binding sites based on the assumption that motif sequences in ChIP-Seq peak regions are highly conserved. Based on such motif sequences featured in the ChIP-Seq regions, we further assumed that the sequences with motifs aligned together are functionally homologous, with higher similarity around the motif sites. Consequently, we inferred the statistical significance for MSAs around binding sites by using motif-anchored gap-free subalignment. In this subalignment, the motif sequence was used as the seed to align to the target sequence at the specific position. The subalignment was completed when both sides of the motif sequence were extended to find the maximal alignment score with the target sequence. The p-value for the specific site was computed as the probability of observing a maximal score at least as high as that of a random sequence. A smaller p-value provides stronger evidence that this motif-anchored query sequence is aligned with such a position in the target sequence.

2.3.5 Training scoring matrices and estimating Karlin-Altschul parameters

In the Karlin-Altschul statistics, the crucial step is to estimate parameters K and \( \lambda \) based on the scoring matrix. As motif shifting could occur in any non-human species, we need to use different scoring matrices and corresponding K and \( \lambda \) for different species in order to test the hypothesis for each species. For each tested species, we extracted MSAs from peak regions with motifs detected in the tested species sequence but no motif shifting occurrence. In such MSAs,
we used all species’ sequences with a motif occurrence to generate consensus sequence. For example, in order to train the score matrix for mouse, the MSA as shown in the Figure 2-5 in the peak region was extracted. The consensus sequence was generated using sequences of human, chimp, rat, and dog. Since there is no motif detected in the cow sequence, the cow sequence was not included. By using this consensus sequence and the tested species’ sequence in all selected peak regions as the training set, the scoring matrix was estimated based on the method described in Chiaromonte’s paper (Chiaromonte. et al., 2002) for each species. The parameter k and λ, then, were estimated based on the trained scoring matrix using the method described in Karlin’s paper (Karlin et al., 1990).

![Figure 2-5: A demonstration for generating the consensus sequence in order to train the score matrix. Gray blocks represent motifs detected on different sequences.](image)

### 2.3.6 Significance test for non-homologous sequences

The significance of the subalignment between the shifted motif-anchored query sequence and the target sequence at the specific position could be inferred from the Karlin-Altschul statistics. To test the null hypothesis that the query sequence is unrelated to the target sequence, the “sequence p-value” (Bailey et al., 1998) was applied. Given a set of sequences with aligned binding sites to human and a sequence with shifting motif, MSAME computes sequence p-values based on the null hypotheses that the sequence anchored with motif shifting are functionally
unrelated, namely non-homologous sequence to the aligned sequences set anchored with binding sites. Or, in another word, the subalignment of target binding site anchored sequence to the query sequence with motif shifting is no better than that to a random sequence with the same length. We performed motif-anchored subalignment between the target and query sequences for each query sequence position by using the motif sequence in the consensus target sequence as the seed and assigned the p-value of such subalignment for each position. Therefore, for a query sequence with length l and target sequence with motif sequence of length m, n = l − m + 1 positions in the query sequence will be aligned. The sequence p-value is calculated by the formula below. The query sequence with ps < 0.05 will be identified as homologous to the target sequences. Query sequences with ps > 0.05 will be categorized as non-homologous sequence.

\[
p_s = 1 - (1 - \min(p_{(1,2,...n)}))^n
\]

2.3.7 The test to classify a motif as either a putative motif shifting or product of local misalignment

The large sequence p-value suggested that the motif shifting in the query sequence was introduced by incorrectly aligning the unrelated sequence to the binding site region. Yet, for a related shifting motif with significant sequence p-value, it could be still possible that the motif shifting is caused by the uncertainty of the local alignment. We attempt to discover if the motif-anchored sequence is aligned better with this specific position than with the aligned binding site position in target sequences set. Therefore, the concepts of background alignment and shifting alignment were introduced. We defined the original alignment, with motif shifting aligned to the original position, as the background alignment (Figure 2-6A). Then we realigned the sequences with shifted motifs to the other sequences by aligning the motifs together and we called this alignment shifting alignment (Figure 2-6B). The same consensus sequence generating strategy
and significant test based on motif-anchored gap-free subalignment were applied to this two alignment. Since we already calculated the p-value for the motif-anchored subalignment for each position, we could acquire the p-values for the background alignment (p<sub>bg</sub>) and shifting alignment (p<sub>sft</sub>) easily. Intuitively, motif shifting was caused by MSA's uncertainty if p<sub>sft</sub> is significant (< 0.05) and p<sub>bg</sub> is not (≥ 0.05), which means the shifting motif-anchored query sequence has strong evidence to relate to the aligned binding site in target sequences. Oppositely, the motif shifting with significant background position p-value (p<sub>bg</sub> < 0.05) will be identified as putative evolutionary motif shifting. In the case that both position p-values are significant, we consider it as putative evolutionary motif shifting as well, because there is no enough information to infer misalignment. Thus, each shifting event was assigned to one of the three different categories based on these two tests.

2.3.8 Unreliability score for MSAs surrounding a binding site

To quantify the quality of MSA, different scores were assigned to three categories of motif shifting. Since we considered non-homologous sequences as the more serious MSA error, we assigned score 1 for motif shifting classified as non-homologous sequences with sequence p-value between 0.05 and 1. Correspondingly, the score of motif shifting caused by local

<table>
<thead>
<tr>
<th>A. Human</th>
<th>CATTGTTACATTGATTTTTTCTATGTAAATTGCT</th>
<th>Human</th>
<th>CATTGTTACATTGATTTTTTCTATGTAAATTGCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimpanzee</td>
<td>CATGTTACATTGATTTTTTCTATGTAAATTGCT</td>
<td>Chimpanzee</td>
<td>CATGTTACATTGATTTTTTCTATGTAAATTGCT</td>
</tr>
<tr>
<td>Gorilla</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
<td>Gorilla</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
</tr>
<tr>
<td>Rhesus</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
<td>Rhesus</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
</tr>
<tr>
<td>Marmoset</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
<td>Marmoset</td>
<td>CATATTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
</tr>
<tr>
<td>Rat</td>
<td>--TTGTATTCTATGTAATATTAT-----</td>
<td>Rat</td>
<td>--TTGTATTCTATGTAATATTAT-----</td>
</tr>
<tr>
<td>Mouse</td>
<td>--ATAATTATATTAT----</td>
<td>Mouse</td>
<td>--ATAATTATATTAT----</td>
</tr>
<tr>
<td>Dog</td>
<td>--TTGTATTCTATGTAATATTAT-----</td>
<td>Dog</td>
<td>--TTGTATTCTATGTAATATTAT-----</td>
</tr>
<tr>
<td>Cow</td>
<td>TATGTTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
<td>Cow</td>
<td>TATGTTGTTACATTGATTTTTTCTATGTAAATTGCT</td>
</tr>
</tbody>
</table>

Figure 2-6: (A) Background alignment: original alignment with motif shifted from positions of motifs in other species (B) Shifting alignment. Alignment after shifting the motif back to make it aligned to motifs in other species.
misalignment with position p-value $p_{bg} \geq 0.05$ was defined as 0.05 to distinguish it from non-
homologous sequences. The score of putative motif shifting, which means no misalignment, was
defined as 0. Meanwhile, to measure the quality of MSA block surrounding the binding site and
all motif shifting related to it, the binding sites in human were applied as the unit for the
measurement purpose. For each binding sites, multiple motif shifting events could be associate to
it from different species. The more species with misalignment, the more unreliable of MSAs
around that binding sites. Based on this fact, we defined the unreliability score for a MSA
surrounding a binding site as the sum of the maximal score of motif shifting in each the species.
Taking alignments with three different type motif shifting in Figure 2-1 as example, the left
alignment of Figure 2-1A will be scored as 0.1, the alignment in Figure 2-1B will be with score
1 and the score for the alignment in Figure 2-1C is 0. Based on this definition of our
unreliability score, the higher the score is, the more unreliable is the MSA block surrounding the
binding site.

2.3.9 Application to genome-wide alignments

MSASE was run on 11-way high-coverage eutherian MSAs generated by MultiZ and
EPO, respectively, to evaluate their alignment quality. Both MSAs were downloaded from their
website and limited to the following species: human, chimp, gorilla, orangutan, macaque,
marmoset, rat, mouse, cow, horse, and dog. The motifs used to conduct evaluation are binding
sites for transcription factors GATA1, CTCF, and NRSF. ChIP-seq peak regions for these three
transcription factors in human in different cell types, generated by the ENCODE project, were
downloaded from the UCSC browser. We extracted MSAs in those human ChIP-seq peak regions
from whole genome-wide MSAs and detected motif sequences occurring in all aligned species
sequences by matching PWMs converted from PFM shown in Table 2-3, 2-4, and 2-5. Both
PFMs of GATA1 and NRSF were downloaded from TRANSFAC with accession number M00128 and M00256. And the PFM of CTCF was acquired from Kim’s paper (Kim et al, 2007).

Table 2-4: The PFM we used for mapping the motif of GATA1

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>10</td>
<td>11</td>
<td>25</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>48</td>
<td>27</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>11</td>
<td>25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>2</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>24</td>
<td>18</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>15</td>
<td>3</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 2-5: The PFM we used for mapping the motif of NRSF

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 2  | 0  | 28 | 0  | 0  | 28 | 0  | 17 | 4  | 1  | 0  | 23 | 0  | 28 | 0  | 10 | 7  | 2  | 4  | 0  | 0  |
| C | 2  | 4  | 27 | 0  | 0  | 27 | 0  | 28 | 28 | 0  | 20 | 0  | 0  | 1  | 26 | 0  | 0  | 17 | 4  | 16 | 22 |
| G | 3  | 3  | 0  | 0  | 28 | 0  | 0  | 0  | 0  | 7  | 2  | 27 | 28 | 0  | 2  | 0  | 28 | 0  | 16 | 8  | 0  |
| T | 21 | 21 | 1  | 0  | 1  | 0  | 0  | 4  | 2  | 0  | 0  | 4  | 0  | 0  | 0  | 1  | 1  | 2  | 2  | 0  | 0  |

Table 2-6: The PFM we used for mapping the motif of CTCF

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 60 | 0  | 46 | 77 | 0  | 0  | 171| 18 | 27 | 229| 0  | 111| 0  | 0  | 0  | 36 | 88 | 0  | 37 | 72 |
| C | 47 | 82 | 41 | 0  | 229| 229| 0  | 112| 144| 0  | 0  | 0  | 0  | 0  | 0  | 0  | 171| 0  | 135| 96 | 74 |
| G | 69 | 40 | 103| 152| 0  | 0  | 30 | 99 | 19 | 0  | 229| 118| 142| 229| 229| 0  | 127| 94 | 32 | 70 |
| T | 55 | 107| 39 | 0  | 0  | 0  | 28 | 0  | 39 | 0  | 0  | 0  | 87 | 0  | 0  | 22 | 14 | 0  | 64 | 13 |

For a negative control, the similar number of random genomic intervals mimic ChIP-seq peak regions with same length distribution were generated and motifs were searched in these regions by using the same threshold as for ChIP-seq peak regions. To analyze the motifs information in each peak regions in different MSAs, we defined a motif conservation score, $S_c$. 
formulized as below to measure how frequently the motifs been detected in these three binding sites through the 11-way eutherian species in each peak region for different MSAs.

$$S_c = \frac{\text{number of sequences with motifs occurrence}}{\text{total number of species aligned in MSAs}}$$

For each binding site identified in human ChIP-seq peak regions, all motif shifting events related to that binding site were reported and classified into three categories based on the criteria discussed above. Finally, we assigned the reliability score for each human binding site as the measurement of the MSAs in those ChIP-seq peak regions.

2.4 Results and Evaluation

Our evaluation method is based on biological criteria, and hence requires the availability of an appropriate validated dataset. However, not many experimental results on the mobility of TFBSs are available. Moreover, most current research on conservation of TFBSs is based on MSAs, so its use could introduce a circularity into our approach. An important departure from this trend is Schmidt’s paper (Schmidt et al. 2012) on the relationship between the expansion of certain retrotransponson families and CTCF binding events, which provides a practical dataset to evaluate our method. However, since our method is based on the occurrence of motif sequences, we first performed an intuitive simulation test to determine the robustness of our program to changes in the threshold cutoff for motif detection. In other words, our approach proceeds in three phases: (1) measure how strongly certain results computed by MSAME depend on the threshold for motif detection, (2) use the “validation” test suite from Schmidt’s paper to evaluate MSAME, and (3) use MSAME to evaluate whole-genome alignments.
2.4.1 Effects of changing the motif detection threshold

We evaluated the robustness of MSAME by changing the threshold for motif detection (the PWM matching threshold) and running the program on both MultiZ and EPO’s MSAs. This performance test is based on the intention that MSAME’s results comparing two MSAs should be robust to changes in the motif detection algorithm and parameters. The noise introduced by false positive motifs should not have significant impact on the evaluation result. Of course, if we lower the threshold for motif detection in all species, the introduction of more identified motifs and more apparent motif shifting may increase the number of events considered for the evaluation, so that the evaluation results could be somewhat different. Even so, the overall qualities of the MSAs are still the same, independent of those conditions.

For this experiment, we retained use of the original threshold for all non-human species, but lowered the threshold for motif detection in the human sequence. This will introduce a number of false positive binding sites and new motif shifting events. We expected the evaluation result to show a significant gain only in putative evolutionary motif shifting, but no significant change in observations of either non-homologous or local-misalignment categories of observed motif shifting, with the same pattern for both MSAs. Any increase of non-homologous and local misalignment identified for new false positive binding sites should occur in the motif shifting events related to closest true positive binding sites. Accordingly, if we continuously lower the threshold, the observed number of motif shifting events classified as non-homologous or local misalignment should be close to constant. Since GATA1 has a relatively short motif and high motif detection rate, we used the GATA1 motif to conduct this experiment on MSAs from both MultiZ and EPO. Figure 2-7 demonstrates that the result matches our expectation. With the threshold decreasing, only the number of motif shifting events classified as putative evolutionary motif shifting climbed significantly. From cutoff 0.9 to 0.8, the increment of motif shifting events
identified as misalignment is linear. Hence, our program’s results are essentially independent of the motif detection threshold and, from the result in Figure 2-7, also independent of input MSAs.

2.4.2 Evaluating MSAME using the datasets in Schmidt’s paper

Subsequently, the performance of MSAME was evaluated using real biological datasets. Schmid et al (2012) identified 5178 ChIP-seq bound regions shared by human, macaque, rat, mouse, and dog. All of the regions were validated by ChIP-seq experiments as true binding sites in all five species. We expected that running our program on this dataset should report few non-homologous sequences in these 5 species. Additionally, the paper listed a group of 130 highly conserved human CTCF binding sites associated with repetitive elements. As shown in Figure 2-8, due to the expansion of repetitive elements, these binding events are surrounded with species- and lineage-specific motifs, which can be taken as true positive evolutionary motif shifting events when evaluating our program. Furthermore, a two-part profile motif of CTCF was discovered in
the five species, which includes a known 20bp motif and another 9bp highly conserved sequence downstream from it by a specific distance (20 or 21bp). This discovery actually, to some extent, supports our assumption that the TFBS is not only conserved in the specific motif sequence but its flanking sequences.

![Diagram of conserved binding sites in five species](image)

Figure 2-8: An example of deeply conserved binding sites in five species (grey blocks) surrounded by lineage-specific (green blocks in rodents) and species-specific (red blocks) motifs presented in Schmid’s paper.

We generated the 5-way shared regions by the method described in Schmidt’s paper, based on the 11-way MSAs mentioned in the Methods section. Since the PWM of CTCF Schmid used in their paper is somewhat different from what we used to detect motifs in MSAs, in the roughly 5000 shared bound regions, MSAME identified 4861 binding sites in human. Around those binding sites, 1380 putative motif shifting events were identified in different species, with 162 of them classified as aligning non-homologous sequences in one of five species, 820 of them classified as results of local misalignment, and the rest of them identified as putative “true” evolutionary motif shifting without misalignment. Since there is no detailed information of true motif shifting in this dataset, we have not enough evidence to verify the motifs classified as results of the local misalignment. Yet, based on the experimental result of ChIP-seq, at least we could assume that all sequences in five species in these 5178 regions should be homologies. Thus, those 162 motif shifting events classified as results from aligning non-homologous sequences are the false positives for identification of non-homologous motif shifting. This implies the
possibility that our method is too stringent for testing non-homologous sequences. Of these 162 cases, 64 and 68 belong to mouse and rat respectively, with another 30 cases coming from dog. The motif shiftings observed in mouse and rat tend to appear together, associated with the same binding sites. Of 132 events identified as non-homologous motif shifti
ng in mouse and rat, 76 are related to 38 shared binding site. Those potential false non-homologous motif shifting events could include cases of true lineage-specific evolutionary motif shifting in rodents. We looked further into these 38 binding sites and the MSA blocks surrounding them, and found that some MSAs are not aligned well enough. If we check MSAs from MultiZ in these regions, the aligned region of mouse and rat could be different. Since the paper identified these 5000 highly conserved regions based on MSAs from PECAN, the MSA error should be considered as well. Therefore, we conclude that the false positive rate for finding non-homologous sequences by our program is not greater than 162/1380 = 11.74%.

We then ran our program on the 130 highly conserved regions associated with retrotransposon-family expansion events. We found that 104 of these 130 regions were detected with CTCF motifs in human by our program. In all, 626 motif shifting events were counted, with 590 of them, in 100 regions, categorized as putative evolutionary motif shifting. Such a high rate of putative shifting matched our expectation as the result of the expansion events of retrotransposons. To verify whether we correctly identified motif shifting, we looked at the regions with expansion patterns shown in the Schmidt’s paper. The B2 expansion specific to the rodent lineage was detected by MSAME. However, we lost the species-specific B2 expansion for mouse and rat individually. The cause is the low sensitivity of our motif detection method. Even so, the result already showed the ability of MSAME to distinguish putative evolutionary motif shifting and misalignment. 21 motif shifting events were classified as non-homologous sequences, 14 of which are from mouse and rat. We checked these 21 alignments and found the same weak alignment as discussed above. Taking region chr1: chr1:40004919-40005339 as an
example, part of the alignment is shown in Figure 2-9. Rat, mouse, and dog were identified as non-homologous sequences by MSAME. When we cross-checked MultiZ alignments, no sequences from mouse, rat, and dog aligned to this region. This disagreement implies that this alignment is potentially unreliable.

Overall, these observations indicate that MSAME is capable of distinguishing MSA errors from putative evolutionary motif shifting, so as to identify the misalignment in MSAs based on motif shifting occurrence. Moreover, our program is robust and independent of motif detection parameters and different MSAs as input. However, the evaluation of MSAME also suggests that our program may overestimate the rate of non-homologous sequences.

2.4.3 MSAME evaluation for MultiZ and EPO MSAs

We then ran MSAME on MSAs generated by MultiZ and EPO. GATA1, CTCF, and NRSF ChIP-seq data sets in human were chosen to evaluate those MSAs based on the motif shifting information in those peak regions. All three transcription factors, especially CTCF, have been shows to be frequently conserved through vertebrate evolution (Mortazavi et al. 2006, Schmidt et al. 2012). Of these three transcription factors, CTCF has the most abundant ChIP-seq peak regions — five times as many as the other two factors. In total, the ChIP-seq peak regions of
these three transcription factors cover around 1.8% of the human genomic. For all three factors, MultiZ has higher alignment coverage in human than does EPO. Approximately 2% of these peak regions have no EPO alignments at all, but only 0.2% of those have no MultiZ alignments. In peak regions with alignments, over 80% of GATA1 peak regions and 45% of CTCF and NRSF regions were identified with motif sequences in human as putative binding sites by MSAME, with binding sites numbering over 31000, 58000, and 16000, respectively. MSAs in those peak regions became our evaluation targets. Thereafter, we counted the number of species covered in the MSAs in each peak region with binding sites and found that MultiZ has more species covered than EPO through all peak regions (Wilcoxon test p-value < 2.2-e16) (see Figure 2-10A). The distributions of conservation scores calculated by formula (3) for each transcription factor in both MSAs are shown in Figure 2-10B. The Wilcoxon test for the scores in all peak regions with binding sites showed that, overall, we could acquire more motif information in MultiZ than in EPO (Wilcoxon test p-value < 2.2-e16). For a negative control, we conducted a similar analysis for random genomic intervals for the three transcription factors. Overall, random genomic intervals have lower motif detection rates: only 40%, 7%, and 19% regions were identified as human binding sites for GATA1, CTCF, and NRSF respectively. The conservation scores for motifs in these random genomic intervals are far less than in ChIP-seq peak regions, with average scores only 0.38 – 0.4 (Wilcoxon test p-value < 2.2-e1.6) (also see Figure 2-10B).
We ran MSAME on 11-way MSAs precomputed by MultiZ and EPO in GATA1, CTCF and NRSF ChIP-seq regions in human. The numbers of binding sites detected with motif shifting events and with unreliability score greater than 0 in both MSAs in each transcription factor peak regions are shown in Figure 2-11A. In all three transcription factor peak regions, MultiZ has a higher percentage of binding sites with unreliable MSAs. Part of the reason is that more motif shifting is detected in MultiZ than EPO, due to the higher species coverage and more motif information acquired, as we showed in Figure 2-10. To further quantify the percentage of unreliable MSAs, we counted the residues covered by unreliable MSAs surrounding the binding sites related to either non-homologous sequences or local misalignment. In total, 7.9% of MultiZ MSAs in all three ChIP-seq regions were evaluated as unreliable, while EPO has 6.7% unreliable MSAs.

Since the results for the three transcription factors are relative consistent, we combined all binding sites together and plotted the distribution of the unreliability scores in both MultiZ and EPO as Figure 2-11B. The distributions of unreliability scores are similar for MSAs from MultiZ and EPO. For both MSAs the medians of the unreliability scores are 1. The Wilcoxon test (p-value = 0.876) also showed that there is no evidence of a significant difference between these two.
distributions. We then plot the percentage of binding sites with unreliable MSAs in GATA1 for four different location categories: Transcription Start Site (TSS) upstream regions, Transcription Termination Sites (TTS) downstream regions, intronic regions, and intergenic regions (Figure 2-11C). MSAs from both MultiZ and EPO follow the same pattern, with a lower rate of misaligned binding sites in TSS and intergenic regions. On the other hand, fold changes between MultiZ and EPO are slightly higher in TSS regions. Thereafter, we looked into the individual motif shifting events to see how many were classified as local misalignment and non-homologous sequence in different MSAs (Figure 2-11D). The percentage based on numbers of motif shifting events suggested MultiZ has a slightly lower rate of non-homologous sequences than EPO. The local misalignment rates are almost the same in both MSAs, consisting of about half of all shifting events. Compared with local misalignment which is possible to be corrected by adjusting gaps, aligning with non-homologous sequences, the completed wrong sequences, is more serious alignment errors. Hence, we are more interested in shiftings caused by non-homologous sequences in different MSAs. We plot the distributions of motif shifting events classified as non-homologous sequences in each species (Figure 2-11E). MultiZ has a lower or equivalent rate of non-homologous MSA errors than EPO in almost all species, except marmoset and dog. The percentage of non-homologous sequences increases dramatically in dog for MultiZ. Both MultiZ and EPO MSAs have a high non-homologous percentage in mouse and rat. Recall that in the performance evaluation discussed earlier, a large proportion of potential false positive homologous sequences were from mouse and rat as well. The high percentage here in mouse and rat could be caused partially by the false positive non-homologous sequences. However, a previous MSA evaluation (Prakash et al. 2007) also showed the same pattern, suggesting a relatively higher percentage of suspicious alignments in mouse and rat than other species.
Figure 2-11: Evaluation results on MSAs pre-computed by MultiZ and EPO in GATA1, CTCF, and NRSF ChIP-seq peak regions. (A) Percentage of binding sites with unreliable MSAs and
2.4.4 Putative evolutionary motif shifting analysis

By ruling out motif shifting classified as either non-homologous or local misalignment, MSASE can identify putative evolutionary motif shifting in other species, relative to human binding sites. Those putative evolutionary motif shifting events reveal potential species-specific binding sites in non-human species. Since the actual number of evolutionary motif shifting events should be independent of MSAs, we expected the numbers of putative evolutionary motif shifting events detected by MSAME for the MSAs to be similar. Our result shown in Figure 2-11D supports this assumption. In three transcription factor ChIP-seq peak regions, in total, the percentages of putative evolutionary motif shifting in both MSAs are very close, both about 55%.

In a further study, we investigated the binding sites with putative evolutionary motif shifting in different region categories, i.e., immediately upstream of the transcription start sites (TSSs), immediately downstream of the transcription termination sites (TTSs), intronic, and intergenic regions (Figure 2-12). Since the putative evolutionary motif shifting is largely independent of the MSA, we only use those identified in MultiZ MSAs for this analysis. Different patterns of occurrence in these four regions were found in GATA1, CTCF, and NRSF ChIP-seq peak regions. Binding sites with putative evolutionary motif shifting in GATA1 ChIP-seq regions have lower abundance in TSSs than in other regions. For CTCF and NRSF peak regions, binding
sites are more prone to be related to a putative evolutionary motif shifting in TSSs than other regions. Since GATA1 is generally bound to TSS proximal regions to regulate gene expression (Yong et al., 2009), the region could be highly conserved through most of vertebrate species. Therefore, fewer motif shifting occur in these regions. This actually could explain the pattern in Figure 2-11C as well. Since CTCF is observed to be bound at various locations relative to the gene, and generally remote from a TSS (Kim et al. 2007), CTCF-bound TSS regions are not as conserved as TSS regions bound by GATA1, which increases the possibility of motif shifting.

To study putative evolutionary motif shifting occurrences in different species, we plot the distribution of putative motif shifting identified in both MultiZ and EPO MSAs in GATA1 peak regions (Figure 2-13A) and the distributions of shifting distances in different species in MultiZ MSAs (Figure 2-13B). The distributions of putative motif shifting in different species are similar in both MSAs (not shown here), suggesting that it is independent of the MSA. The result shows that the more evolutionarily distant the species is, the more motifs are prone to shift away from the binding sites in human. However, mouse and rat are two exceptions. Compared with cow, horse, dog, and even marmoset, mouse and rat have a lower percentage of motif shifting occurrence and shorter motif shifting distances. This may be due to the high percentage of motif shifting identified as either non-homologous or local misalignment in mouse and rat. That both
mouse and rate have shorter shifting distances could be the result of linear-specific binding sites in rodent. Nevertheless, it cannot explain why the distances are shorter than other species, an observation worthy of further study.

Figure 2-13: Putative evolutionary motif shifting in different species. (A) Distribution of
Chapter 3

Discussion and Future work

3.1 Discussion

MSA evaluation is not a problem that could be solved easily by defining “good” or “bad” for different alignments. Even so, we could apply a criterion to measure the reliability of an MSA for downstream analyses. In this paper, we proposed a new method, called MSAME, the first MSA evaluation method based on experimental data from the ENCODE project. Instead of using simulations that rely on evolutionary models, we defined a biological criteria inferred from different scenarios of occurrences of motif shifting in ChIP-seq peak regions, and evaluated MSAs from such a functional perspective. By using motif-anchored realignment and a significance test, unreliable MSAs surrounding binding sites with motif shifting events were identified.

Evaluation of performance showed that MSAME could identify unreliable MSAs and putative evolutionary motif shifting efficiently. Nevertheless, the evaluation also suggested that our program may overestimate the percentage of non-homologous sequences. One possible reason is the gap-free realignment we used to infer the p-value for the null hypothesis of unrelated sequences. We assumed that the TFBSs in ChIP-seq peak regions should be highly conserved, not only conserved in motif sequences but also conserved in its flanking sequences. Even so, non-coding regions are far less conserved than coding regions and, generally, with a number of
insertions and deletions introduced through evolution. A significant gap-free alignment may be a too stringent condition to be satisfied here in some cases. However, such a stringent criterion provided a guarantee for the low false negative (false reliable MSAs), which is good for downstream analysis. Especially if our purpose is to identify putative evolutionary motif shifting, this high false positive rate in unreliable MSAs will keep a low false positive putative evolutionary motif shifting identification.

Unfortunately, we used only one biological dataset for evaluating false positive non-homologous sequences, and presented no evaluation for local misalignment. This was because there is limited research published about binding site mobility or motif turnover, so little real biological data about true motif shifting is available to evaluate our method. For now, since not much empirical data is available, we have no standards to adjust the cutoff of p-value for both non-homologous and local misalignment tests, so we used 0.05 based on “intuition”. However, this parameter should be adjusted based on rigorous evaluations for our method, which we hope will become possible when more ChIP-seq data is published for more species.

One limitation of the evaluation method based on biological data is that inadequate biological data will make the performance of such methods suffer. As for our method, the limited knowledge of motif PWMs or consensus sequences decreases the range of MSAs that our method can evaluate. Since our method relies on motifs identified in human ChIP-seq regions to define motif shifting, only regions with detected motifs will be evaluated. The low accuracy of PWMs leads to a low rate of regions detected with motifs. Additionally, although our method is robust to the noise introduced by false positive motifs and largely independent of parameters for detecting motifs, it is still highly dependent on the occurrence of motifs in all species in MSAs. The accuracy of a PWM for a motif determines how many motifs we could detect in the sequences. In a MSA, our method could only evaluate aligned sequences with motif occurrences. For those sequences without the motif identification, we have no biological standard for our tests. For
MSAs with very few motif sequences, the evaluation is difficult due to the lack of motif shifting information. Moreover, the introduction of false positive motifs in non-conserved regions may cause higher false positive identification of non-homologous sequences. Therefore, with more PWM research, will undoubtedly be published in the future, the performance of our method could be improved as well.

Currently, due to limited species available with ChIP-seq data, we only used human ChIP-seq data, making our definition of motif shifting quite human-centric. In the future, with more ChIP-seq data released in different species, it will become reasonable to consider combining motif information from different species. The motif shifting definition should be extended to multiple species as well. At that point, more MSAs will be evaluated by this method, giving better performance.

3.2 Future work

Our method is not only an evaluation method; it could to be extended for different biological applications. One additional benefit of our method is that, as we discussed in the Results section of this manuscript, we could use it to detect putative evolutionary motif shifting or species-specific motifs, so that the conserved motif detection won’t depend on aligned positions in MSAs anymore. Thereafter, difference among species in positioning of motifs could be predicted as functional variance. Furthermore, our method could be extended to improve MSA quality. Since our method is based on motif-anchored realignment, and the test result for local misalignment is to report the best location where the shifted motif should be aligned, the method could be extended to be an “anchored realigner” for improving an MSA’s quality.
REFERENCES CITED

Altschul. 1997. Evaluating the statistical significance of multiple distinct local alignments, In *Theoretical and computational methods in genome research* (pp. 1-14). Springer US.


Margulies E.H., Cooper G., Asimenos G., Thoma D.J., Dewey CN, et al. 2007. Analysis of deep mammalian sequence alignments and constraint prediction for 1% of the human genome. *Genome Res.* 17: 760-774


Wise MJ. 2010 No so HoT - heads or tails is not able to reliably compare multiple sequence alignments. *Cladistics* 26:438-443