Specificity and methylation sensitivity analysis of HoxB13

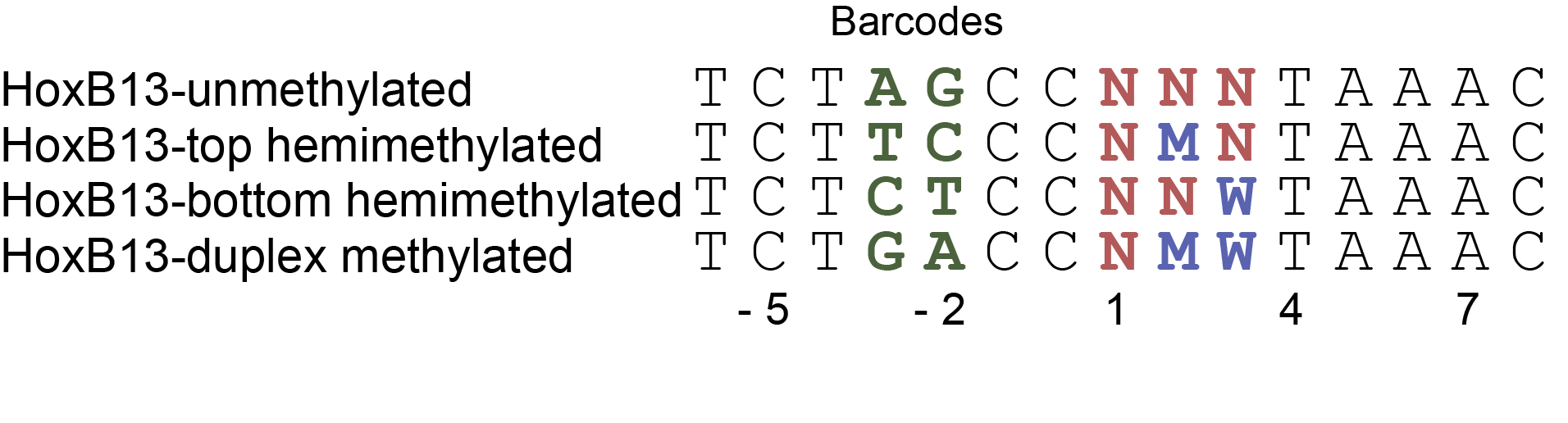
Zheng Zuo

Sep 16, 2020

Table of Contents

## Introduction

HoxB13, is a homeobox protein recently found to process methylation sensitivity and two distinct motifs, i.e., TCGTAAA and CAATAAA. To quantify this, I designed following randozmied dsDNA libraries covering the unmethylated, top hemimethylated, bottom hemimethylated, duplex methylated binding sites at positions 2 and 3 respectively by chemical synthesis, thus we can assay the strand-specific methyl-C contribution to the binding affinity of each site.



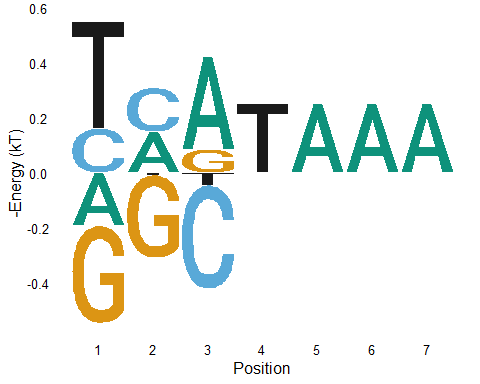
## Importing and preoprocessing data

load("../data/HoxB13.RData")  
  
(HoxB13 <- HoxB13 %>%  
 dplyr::mutate("Bound/Unbound" = Bound/Unbound,  
 Energy = -log(Bound/Unbound)))

## # A tibble: 100 x 6  
## Sequence Property Bound Unbound `Bound/Unbound` Energy  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 CCATAAA top Hemimethylated 119049 3949 30.1 -3.41   
## 2 TCGTAAA duplex Methylated 246730 16406 15.0 -2.71   
## 3 TCATAAA top Hemimethylated 113243 8850 12.8 -2.55   
## 4 TCGTAAA top Hemimethylated 157847 12818 12.3 -2.51   
## 5 TCGTAAA unmethylated 66312 12149 5.46 -1.70   
## 6 ACATAAA top Hemimethylated 42781 8102 5.28 -1.66   
## 7 TCGTAAA bottom Hemimethylated 40646 8448 4.81 -1.57   
## 8 TCATAAA unmethylated 15940 9614 1.66 -0.506  
## 9 CCGTAAA duplex Methylated 23539 15208 1.55 -0.437  
## 10 CAATAAA unmethylated 41067 27166 1.51 -0.413  
## # ... with 90 more rows

## Building specificity model

HoxB13 %>%  
 dplyr::filter(Property == "unmethylated") %>%  
 TFCookbook::buildEnergyModel() %>%  
 TFCookbook::getEnergyMatrix() %>%  
 TFCookbook::addAnchorMatrix(anchor = "TAAA", position = 4, height = 0.25) %>%  
 TFCookbook::plotEnergyLogo() +  
 scale\_y\_continuous(breaks = seq(-0.8, 0.8, 0.2))



## Building methylation sensitivity model

It is easy to build methylation effect model by comparison of each methylated ste with its unmethylated counterpart

HoxB13.unmethylated <- subset(HoxB13, Property == "unmethylated")  
HoxB13.methylated <- subset(HoxB13, Property %in% c("top Hemimethylated", "bottom Hemimethylated", "duplex Methylated"))  
  
(HoxB13.paired <- left\_join(HoxB13.methylated, HoxB13.unmethylated, by = "Sequence") %>%  
 dplyr::mutate(Energy = Energy.x - Energy.y) %>%  
 dplyr::select(Sequence,  
 Property = Property.x，  
 Energy))

## # A tibble: 36 x 3  
## Sequence Property Energy  
## <chr> <chr> <dbl>  
## 1 CCATAAA top Hemimethylated -3.08   
## 2 TCGTAAA duplex Methylated -1.01   
## 3 TCATAAA top Hemimethylated -2.04   
## 4 TCGTAAA top Hemimethylated -0.814  
## 5 ACATAAA top Hemimethylated -2.86   
## 6 TCGTAAA bottom Hemimethylated 0.126  
## 7 CCGTAAA duplex Methylated -1.25   
## 8 CCGTAAA top Hemimethylated -0.977  
## 9 TAGTAAA bottom Hemimethylated -0.632  
## 10 ACGTAAA top Hemimethylated -0.902  
## # ... with 26 more rows

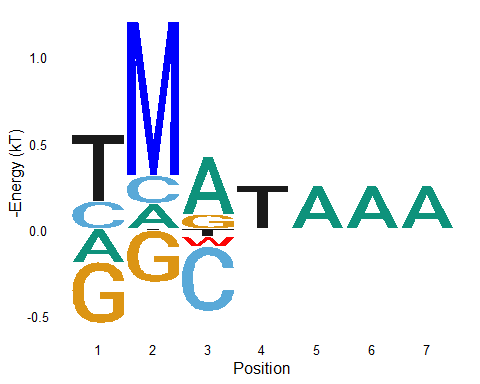
Overall, we can create a compositive model to include both specificity and methylation effect

HoxB13.MethylModel <- HoxB13.paired %>%  
 dplyr::mutate(Sequence = dplyr::case\_when(  
 Property == "top Hemimethylated" ~ stringi::stri\_sub\_replace(Sequence, from=2,to=2, replacement = "M"),  
 Property == "bottom Hemimethylated" ~ stringi::stri\_sub\_replace(Sequence, from=3,to=3, replacement = "W"),  
 Property == "duplex Methylated" ~ stringi::stri\_sub\_replace(Sequence, from=2,to=3, replacement = "MW")  
 )  
 ) %>%  
 TFCookbook::buildMethylationModel()

## Warning: Expected 14 pieces. Additional pieces discarded in 36 rows [1, 2, 3, 4,  
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

HoxB13 %>%  
 dplyr::filter(Property == "unmethylated") %>%  
 TFCookbook::buildEnergyModel() %>%  
 TFCookbook::getEnergyMatrix() %>%  
 TFCookbook::addAnchorMatrix(anchor = "TAAA", position = 4, height = 0.25) %>%  
 TFCookbook::addMethylMatrix(MethylModel = HoxB13.MethylModel, encoding = "(3+2)L+1") %>%  
 TFCookbook::plotEnergyLogo()

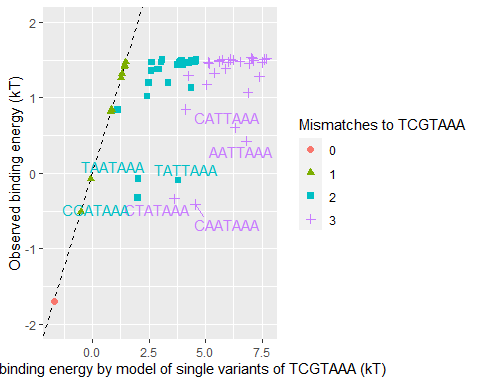
## Warning: Expected 21 pieces. Additional pieces discarded in 64 rows [1, 2, 3, 4,  
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].



## The existence of secondary motif and the necessity to use composite “Min” model to predict binding energy

Since among all unmethylated sequences, TCGTAAA has the lowest binding energy and hightest affinity, we can build a specificity model based on all those single variants to TCGTAAA as below. Using this model to predict the binding energy for all other sites and compare them with experimentally observed values, it is clear that small fraction of sequences, e.g., CAATAAA and CTATAAA, deviate significantly from the diagnoal lines, meaning that they are not properly represented in the energy model made by TCGTAAA and its single variants. The most intuitive explanation is that there exists another structural conformation for the recognition of DNA sequence CAATAAA and its related variants. We can build another model from CAAATAAA and its single variants, called HoxB13.CAA.Model as below. However, this “CAA” model still wouldn’t explain all the observed values, particularly for sequence like TCGTAAA. Alternatively, if for each sequence we take the minimal value predicted by “TCG” and “CAA” models as the composite predicted value and compare it with the observed number, they match very well, so it is fair to say the TCG and CAA models are both required for proper understanding the specificity behavior of HoxB13 to its unmethylated sites.

HoxB13.TCG.Model <- HoxB13 %>%  
 dplyr::filter(Property == "unmethylated") %>%  
 TFCookbook::selectVariants(reference = "TCGTAAA", maxMismatches = 1) %>%  
 TFCookbook::buildEnergyModel()  
  
HoxB13 %>%  
 dplyr::filter(Property == "unmethylated") %>%  
 mutate(Predicted.Energy = TFCookbook::predictEnergy(Sequence, HoxB13.TCG.Model),  
 Mismatch.TCG = as.factor(TFCookbook::countMismatch(Sequence, "TCGTAAA"))) %>%  
 ggplot(aes(x = Predicted.Energy, y = Energy, label = Sequence, color = Mismatch.TCG, shape = Mismatch.TCG)) +  
 geom\_abline(slop = 1, linetype = "dashed") +  
 geom\_point(size = 2) +  
 geom\_text\_repel(data = function(x) filter(x, ((Predicted.Energy - Energy) > 2) & (Energy < 0.8)), show.legend = FALSE) +  
 xlab("Predicted binding energy by model of single variants of TCGTAAA (kT)") +  
 ylab("Observed binding energy (kT)") + ylim(-2, 2) +  
 labs(color = "Mismatches to TCGTAAA", shape = "Mismatches to TCGTAAA")



HoxB13.CAA.Model <- HoxB13 %>%  
 dplyr::filter(Property == "unmethylated") %>%  
 TFCookbook::selectVariants(reference = "CAATAAA", maxMismatches = 1) %>%  
 TFCookbook::buildEnergyModel()  
  
HoxB13 %<>%  
 dplyr::filter(Property == "unmethylated") %>%  
 dplyr::mutate(Mismatch.TCG = as.factor(TFCookbook::countMismatch(Sequence, "TCGTAAA")),  
 Mismatch.CAA = as.factor(TFCookbook::countMismatch(Sequence, "CAATAAA")),  
 Predicted.Energy.TCG = TFCookbook::predictEnergy(Sequence, HoxB13.TCG.Model),  
 Predicted.Energy.CAA = TFCookbook::predictEnergy(Sequence, HoxB13.CAA.Model)) %>%  
 dplyr::mutate(Predicted.Energy.min = pmin(Predicted.Energy.TCG, Predicted.Energy.CAA))  
  
HoxB13 %>%  
 ggplot(aes(x=Predicted.Energy.TCG, y=Energy, label=Sequence, color=Mismatch.TCG, shape = Mismatch.CAA)) +  
 geom\_abline(slop = 1, linetype = "dashed") +  
 geom\_point(size = 2, show.legend = FALSE) +  
 geom\_text\_repel(data = function(x) filter(x, ((Predicted.Energy.TCG - Energy) > 2) & (Energy < 0.8)), show.legend = FALSE) +  
 xlab("Predicted binding energy\n by model of single variants of TCGTAAA (kT)") +  
 ylab("Observed binding energy (kT)") + ylim(-2, 2) -> plot.TCG  
   
HoxB13 %>%  
 ggplot(aes(x=Predicted.Energy.CAA, y=Energy, label=Sequence, color=Mismatch.TCG, shape = Mismatch.CAA)) +  
 geom\_abline(slop = 1, linetype = "dashed") +  
 geom\_point(size = 2, show.legend = FALSE) +  
 geom\_text\_repel(data = function(x) filter(x, ((Predicted.Energy.CAA - Energy) > 1.5) & (Energy < 0.8)), show.legend = FALSE) +  
 xlab("Predicted binding energy\n by model of single variants of CAATAAA (kT)") +  
 ylab("Observed binding energy (kT)") + ylim(-2, 2) -> plot.CAA  
  
HoxB13 %>%  
 ggplot(aes(x = Predicted.Energy.min, y = Energy, label = substr(Sequence, 1,3), color = Mismatch.TCG, shape = Mismatch.CAA)) +  
 geom\_abline(slop = 1, linetype = "dashed") +  
 geom\_point(size = 2) +  
 geom\_text\_repel(data=function(x) filter(x, (abs(Predicted.Energy.min - Energy) > -0.1) & (Energy < 1)), show.legend = FALSE) +  
 xlab("Predicted binding energy\n by minimal value of TCGTAAA and CAATAAA models (kT)") +  
 ylab("Observed binding energy (kT)") + ylim(-2, 2) +  
 labs(color = "Mismatches to TCGTAAA", shape = "Mismatches to CAATAAA") -> plot.min  
  
cowplot::plot\_grid(plot.TCG, plot.CAA, plot.min, align = "h", ncol = 3, rel\_widths = c(1,1,1.45))

