Detect Positive Selection in Type IIL Restriction Enzyme Using PAML

Yiyi Pu
Hangzhou, Zhejiang, China

Bachelor of Agronomy, Zhejiang University, 2015

A Thesis presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Master of Science

Department of Biology

University of Virginia
December, 2017

Abstract

The role of positive selection in molecular evolution has long been debated since the proposal of the neutral theory. With the explosive growth of genomic sequence data and aided by statistical methods such as Phylogenetic Analysis by Maximum Likelihood (PAML), recent years witnessed a flurry of reports of positive selection in protein-coding genes. However, with a few exceptions, the reported positively selected sites generally lacked functional data to support them. Restriction enzymes in bacteria are under constant selective pressure from the bacteria-phage arms race, and in some enzymes amino acids responsible for the specificity changes are known. Here I tested the performance of PAML using the functionally well characterized type IIL restriction enzymes as the model system. The result showed that PAML was highly consistent in detecting positive selection in the sequence data. However, in terms of identifying positively selected sites, PAML was very sensitive to the sampling bias in the dataset and had a high false negative rate and also possibly a high false positive rate. The result suggests positively selected sites identified by PAML should be treated with caution and validated by functional studies.

Acknowledgements

I would like to give my first thank to my parents who gave me warm support in so many aspects during the two years. I also want to give a special thank to my ‘host family’ here. When I was together with them, I felt settled and at home. For academic part, first I want to thank my mentor Martin Wu who accepted me as his student two years ago. During the two years, he is always really patient and helpful especially when the project was in its bottleneck. Second, I would like to thank previous and current Wu lab members, Tiantian Ren, Yingnan Gao, Maurice Wong, for all their help and suggestions. Third, I would like to thank my committee members, Robert Cox and Alan Bergland, for the time they spent on committee meetings and reviewing my materials and also for useful suggestions they proposed. At last, I would like to thank the previous myself for not giving up during hard times. This thesis is a present to me in ‘my year’.

Introduction

Positive selection is defined as the process where variants in a population provide higher fitness and thus are favored by selection. The importance of positive selection has long been accepted at the morphological trait level. However, whether it is important at the molecular level is still debated. The neutral theory proposed that instead of natural selection, most evolutionary changes at the molecular level are caused by random fixation of neutral mutations (Kimura, 1983). Although it does not totally exclude positive selection, it does assume that positive selection is rare and thus does not play a major role in evolution at the molecular level. Ever since the proposal of neutral theory, there has been ongoing debate about the role of positive selection in molecular evolution.

Recent or ongoing positive selection

The advent of molecular biology made it possible to study molecular evolution using DNA and protein sequences. With accumulated sequence data, statistical methods have been developed for studying evolution at the molecular level. For recent or ongoing positive selection, mainly three tests are used by researchers. The first is the McDonald-Kreitman (MK) test that compares the nonsynonymous/synonymous substitution rate ratio within species (pN/pS) and between species (dN/dS) (McDonald & Kreitman, 1991). Because positive selection has a larger impact on increasing dN than pN, dN/dS significantly higher than pN/pS indicates positive selection. The second is linkage disequilibrium (LD). LD refers to non-random association of alleles at different loci. Under neutral evolution, recombination causes LD around a new allele to decay substantially because it takes the allele a long time to reach high frequency. However, positive selection raises allele frequency too rapidly for recombination to break up the LD. Based on this principle, long-range LD is used as an indicator of positive selection (Sabeti et al., 2002). The third method is Fixation Index (F_{ST}) test (Lewontin & Krakauer, 1973). Recent positive selection can generate differentiation between populations. F_{ST} is a measure of population difference and significantly different distribution of F_{ST} among loci from the neutral expectation can be used to infer positive selection (Akey et al., 2002).

Using above methods, researchers started to find evidence of positive selection at the molecular level. Researchers not only can conduct MK test but also can calculate α (1 - dSpN/dNpS), which indicates the proportion of adaptive substitutions. Using the MK test, studies on various Drosophila species found substantial proportions of amino acid substitutions driven by positive selection (~45% - Smith & Eyre-Walker, 2002; ~25% - Bierne & Eyre-Walker, 2004, ~54% - Begun et al., 2007). Besides Drosophila, another study surveyed animals from all phyla (e.g. insects, molluscs, annelids, echinoderms, reptiles, birds, and mammals) and concluded that for the majority of animals, more than 50% of amino acid substitutions were driven by positive selection (Galtier N, 2016). Similar result (more than 50% adaptive substitutions) was also reported in enteric bacteria (Charlesworth & Eyre-Walker, 2006). After reviewing such studies, Hahn proposed a ‘Selection Theory’ in which he argued that positive selection is more prevalent and should serve as the null model for population genetics studies (Hahn, 2008). Besides MK test, positive selection was also detected by the other two methods. By long-range LD, widespread signals for recent positive selection were found in the human genome (Voight et al., 2006; Sabeti et al., 2007; Williamson et al., 2007). Positively selected genes in various organisms (e.g., fishes, plants and human) were also identified by high F_{ST} by various studies (Nielsen et al., 2009; Hohenlohe et al., 2010; Strasburg et al., 2009; Namroud et al., 2008; Akey et al., 2002; Barreiro et al., 2008; Myles et al., 2008).

However, analyses in several model systems found no evidence of positive selection. No positive selection was detected in 330 human genes by the MK test comparing human, chimpanzees and Old World monkeys (Zhang & Li, 2005). Similarly, no positive selection was found in Arabidopsis or yeast (Foxe et al., 2008; Doniger et al., 2008). In addition these population genetics-based methods often do not have power to determine whether departure from the neutral model is due to positive selection or changes in population size. For example, for MK test, Eyre-Walker found that increase of effective population size can produce artificial evidence of positive selection when there were slightly deleterious amino acid substitutions (Eyre-Walker, 2002). For the other two methods, high false-positive problem has been reported. Akey reviewed several genome-wide positive selection studies in humans and found only 14.1% positive selected regions were identified by multiple studies (Akey, 2009).

Past positive selection

For past positive selection, the dN/dS (nonsynonymous/synonymous substitution rate ratio, also called ω) is the most widely used metric for detecting positive selection. Assuming selection mostly acts only on nonsynonymous substitutions, ω can be used as a measure for selective pressure. In principle, ω can be binned into three categories: 1) ω close to 1 - neutral substitution. In this case nonsynonymous and synonymous substitutions are fixed at the same rate, which indicates no selection. 2) ω < 1 - purifying selection. In this case nonsynonymous substitutions are deleterious and thus are removed by selection. ω close to 0 indicates strong purifying selection that removes almost all nonsynonymous substitutions. 3) ω > 1 - positive selection. In this case, nonsynonymous substitutions provide selective benefits and are fixed at a higher rate than synonymous ones (Yang & Bielawski, 2000).

Traditional methods calculate an average dN/dS ratio for the entire protein sequence. This is highly conservative, because most sites of the protein may be under strong purifying selection (with dN/dS close to 0) and therefore the average dN/dS will be well under 1 even when some sites are under positive selection. Yang's group developed a ML method (PAML) that allowed ω to vary among sites. Two steps are involved: 1) a likelihood-ratio test (LRT) between the null model (ω <= 1) and the alternative positive selection model (ω <= 1 & ω > 1) is conducted. If LRT suggests positive selection, then 2) sites with high posterior probability to have ω > 1 are identified using a Bayes approach (Yang & Bielawski, 2000; Yang, 2007). Later, a branch-site model was developed to detect positive selection that only affects sites in specific lineages by allowing ω to vary both among sites and among lineages (Goldman & Yang, 1994).

Using dN/dS, various genes have been studied and showed evidence of positive selection (Ford, 2002). Among those genes, one classical example was the Major Histocompatibility Complex (MHC) gene in vertebrates (Hughes & Yeager, 1998) whose protein products present antigen peptides to T cells in the immune system. Hughes and Yeager tested the hypothesis that positive selection acting on peptide-binding region (PBR) maintains the polymorphisms of MHC by calculating dN and dS of PBR and non-PBR. They found significant higher dN relative to dS only in PBR, indicative of positive selection. Besides MHC, many immune-related genes have been shown to be under positive selection. Positively selected sites were identified in various members of Toll-like receptor gene family in mammals (TLR2 - Tschirren et al., 2011; TLR9 - Park et al., 2010; TLR22 – Sundaram et al., 2012), as well as in the CC chemokine receptor genes of mammalian immune system (Metzger & Thomas, 2010). Ford surveyed over 100 published studies and found 119 genes with statistical evidence of positive selection and found the largest group (47 genes) was involved in either host defense or parasite response (Ford, 2002). Similarly, a study that surveyed all genes in the Drosophila genomes also found that immune response genes contained extremely high proportion of positively selected codons (Heger & Ponting, 2007). This makes sense because immune system genes are thought to be under strong positive selection due to ongoing arms races between hosts and pathogens (Heger & Ponting, 2007; Aguileta et al., 2009; Ford, 2002).

Despite all these findings, multiple studies have also challenged PAML. False-positive problems have been reported by several papers. Suzuki and Nei found that PAML produced an unacceptably high false-positive rate in a simulation study (Suzuki & Nei, 2002). Hughes and Friedman showed that due to the stochastic nature of substitutions, dN/dS > 1 can happen just by chance and the chance was higher for short branches (Hughes & Friedman, 2008). By calculating the number of nonsynonymous and synonymous substitutions, they also found that sites identified by the branch-site model were largely due to lack of synonymous substitutions (Hughes & Friedman, 2008). One of the more convincing cases against PAML was presented by Yokoyama et al. using dim-light vision proteins in vertebrates as their model system (Yokoyama et al., 2008). Using phylogenetic methods, they inferred ancestral states of the protein and then engineered 11 ancestral proteins by mutagenesis. After comparing the phenotypes of ancestral proteins and extant proteins, they identified 15 adaptive changes at 12 amino acid sites. However, PAML identified 8 totally different sites under positive selection (with $\omega > 1$) but none of those sites were associated with any functional change. Disagreement between the experimental results and PAML predictions demonstrated that detecting positive selection purely by statistical methods can be problematic. Using a similar approach, Zhuang et al. studied odorant receptor genes in primates and found little overlap between sites predicted by PAML and those with experimental support (Zhuang et al., 2009).

PAML has been widely used and become influential to our understanding of the role of the positive selection in molecular evolution. However, there is still an ongoing debate about whether findings from PAML analysis are reliable and biologically meaningful. Due to the lack of model systems with functional data, PAML's performance was mostly evaluated by simulation studies (Anisimova et al., 2001, 2002; Wong et al., 2004; Yang & dos Reis, 2011). There is a clear need for more studies in real systems to test how well PAML works. Using type IIL restriction enzymes in bacteria as the model system, I will try to answer two questions: 1) Can PAML detect positive selection? 2) Does PAML correctly identify positively selected sites?

The model system

Bacteria in the natural environment are constantly attacked by phages, which outnumber the bacterial cells by a ratio of 10:1. In response, bacteria have evolved immune systems such as the restriction-modification system (RM system) to protect against phage infections, leading to a pronounced evolutionary arms race (Stern & Sorek, 2010). All RM systems consist of two functional modules, a methyltransferase to protect the host DNA by methylating specific nucleotide bases within the recognition sequence and an endonuclease to cleave unprotected foreign DNA at or near the recognition sequence (i.e., restriction site). The enormous selective pressure from the rapid turnover and evolution of phage particles has driven rapid evolution and diversification of the RM genes, particularly those involved in recognizing the restriction sites (Sharp et al., 1992; Murray et al., 1993; Zheng et al., 2004). There are 4,596 known restriction enzymes from more than 900 species that recognize 729 restriction sites (http://rebase.neb.com/rebase/rebase.html).

Unlike most genes studied for positive selection, the precise functions including the recognition sequence are known for most RM genes (Roberts et al., 2015). The type II RM system is the most common type with high specificity in both methylation and cleavage. For most type II RM systems, the methyltransferase and endonuclease are encoded by two separate genes and also act independently. However, both proteins need to recognize the same target DNA sequence (i.e., restriction site) for normal function. Specificity change in merely one module or non-synchronous change in two modules will cause the host DNA to be cleaved by its own endonuclease. For this reason, there is strong constraint for evolutionary changes for type II RM, since synchronous changes are rare. However, in one family of type II RM proteins (type IIL REs), the two functional modules are merged into one protein and they share one target DNA recognition domain (TRD) (Callahan et al., 2016). As a result, the TRD of type IIL REs should be more tolerant to specificity changes and amenable to positive selection.

In support of this idea, evolutionary changes of type IIL RE specificity are frequent and labile. Among 21 type IIL REs compared in a study, each had its unique specificity (Morgan & Luyten, 2009). With a few exceptions, type IIL REs recognize 6-base restriction sites. Most bases in the recognition sequence are free to change with the exception of the 5th base, which is highly conserved. The molecular basis of sequence recognition in type IIL RE has been elucidated. Amino acids that correlated with the specificity changes were identified and 7 of them (corresponding to positions 645, 751, 773, 774, 806, 808, 810 in MmeI, Figure 1) were confirmed to cause specificity change at base 2, 3, 4 and 6 by mutagenesis experiment and the structure of enzyme-DNA complex (Morgan & Luyten, 2009; Callahan et al., 2016). Because those specificity changes correspond to functional changes in nature, the 7 sites are expected to be under positive selection to match the rapid evolution of the restriction sites in phages. As such, type IIL REs represent an excellent model system for studying adaptive molecular evolution.

Materials and Methods

Sequence datasets

The nucleotide, protein and recognition sequences of 47 type IIL REs (referred to as ‘original sequences’ hereafter) were downloaded from REBASE. Homologs of the 47 original sequences were downloaded from NCBI as follows. NCBI tblastn search with 6 representative original sequences (the seed sequences) as the query was used to identify DNA sequences of type IIL RE homologs (E-value cutoff 1e-30). However, since there is no option to download matched CDS (coding DNA sequence) directly from NCBI, I chose to download genomes of all bacterial species that had a match in the blastp search. Because downloading all genomes for each species (~1,300 species, ~63,000 genomes in total) is impractical, and also because many homologs in different strain genomes of the same species are identical, I chose to download 2 genomes for each species. CDS were extracted from the download genomes and searched against by local tblastn to retrieve type IIL RE homologs. After excluding identical sequences, partial sequences and homologs that formed an outgroup, 274 homologs were retained. The entire dataset consisted of 321 sequences (47 original sequences and 274 homologs).

Simulation studies showed that sequence divergence affected the accuracy and power of PAML analysis (Anisimova et al., 2001, 2002; Wong et al., 2004; Yang & dos Reis, 2011). Accuracy was defined as the probability that a site predicted to be under positive selection was truly under positive selection, and power was defined as the probability that a site truly under positive selection was predicted to be under positive selection. To test the effect of sequence divergence on PAML analysis, I downloaded additional sequences that are close to the 6 seed sequences. Homologs were downloaded as above with one difference: all genomes of species in the top 50 hits of the blastp search were downloaded. I selected four subtrees that contained the seed sequences to assess the effect of sequence divergence on PAML analysis (Figure 2). Within each subtree, different levels of sequence divergence were achieved by iteratively extracting smaller subclades using the original sequences as anchors till about 10 sequences were left. Sequence divergence was measured in two different ways. The max dS measures the synonymous substitution rate (dS) between the two most distantly related sequences in the dataset, and the average dS is the average synonymous substitution rate between all pairs of sequences in the dataset.

To further expand the range of sequence divergence for PAML analysis, I also compiled a dataset consisting of homologs from a single species, *Corynebacterium diphtheriae*. This species was chosen because in the original dataset, there was one pair of highly similar sequences from this species with different specificities, suggesting possible recent positive selection. I downloaded 180 *C. diphtheriae* genomes from NCBI from which I identified 28 unique homolog sequences.

Phylogenetic analysis

For all analyses in this study, protein sequences were aligned using MAFFT (version 7.205, Katoh & Standley, 2013) and the alignment was trimmed by ZORRO (Wu et al., 2012) with the cutoff 4 to remove low-quality aligned columns. The protein alignments were used to guide the alignment of corresponding nucleotide sequences using an in-house perl script. RAxML (version 8.2.4, Stamatakis, 2014) was used in all phylogeny reconstruction. Substitution models used for protein and nucleotide alignment were PROTCATWAG and GTRCAT respectively. The topology of phylogenetic tree was based on the trimmed protein alignment and the branch lengths estimated using the nucleotide alignment were multiplied by 3 to get codon-based branch lengths for the following dN/dS ratio analyses.

Ancestral state/sequence reconstruction

Using the phylogeny of the 47 original sequences, the ancestral states of recognition site were reconstructed by parsimony using Mesquite (version 3.31, Maddison WP & DR Maddison, 2017). The ancestral sequences of the enzymes were reconstructed by maximum likelihood method using CODEML in PAML (version 4.9e, Yang, 2007).

PAML analysis

Two models of CODEML in PAML (version 4.9e, Yang, 2007) were used. The site model that allowed ω ratio to vary across all codons was used unless otherwise noted. I compared the recommended model pair M7 (beta) and M8 (beta&ω). M7 assumes a beta distribution (in the interval [0, 1]) for all ω among sites and serves as the null model. M8 adds one more class of ω that allows ω to be higher than 1 and serves as the alternative model. The likelihood ratio test was conducted for the pair of models first and if the difference was statistically significant, a Bayesian approach was then used to calculate the posterior probability of $\omega > 1$ for each site and sites with a probability higher than 0.95 were identified as positively selected sites. To avoid likelihood being trapped in a local optimum, for each analysis I tried two initial ω values (0.4 and 1.5) as the starting points and only included the result with the higher likelihood. In all analyses except one case (the whole dataset), starting with different initial values always resulted in convergence to the same likelihood.

Because the 6th base in the recognition sequence is partially conserved (G or C), the branch-site model was applied to detect positive selection that might only act on a few branches (branches involving changes between G and C). The branch-site model was only applied on the original sequences because functional information was required to bin the branches into two different categories. I applied the branch-site model on the 47 original sequences and the 10 original sequences from subtree 2.

Bootstrap analysis

To investigate the consistency of PAML, I conducted a bootstrap analysis with 100 pseudoreplicates. I used subclade 2 dataset of subtree 2 for bootstrap analysis because PAML identified the largest number of sites under positive selection in this dataset. For each pseudoreplication, I randomly selected 50 sequences without replacement from the 82 sequences of the subclade 2 and carried out a PAML analysis as described above. The bootstrap support of a site was calculated as the number of times it was identified as positively selected sites in the 100 pseudoreplicates.

Results

Evolution of type IIL RE specificity

As expected, the evolution of recognition sequences of type IIL REs was rapid (Figure 3). Among the 6 bases of the recognition site, only the 5th base was highly conserved (A). The 6th base changed with some constraint (G/C) while the remaining four bases all changed frequently. Using parsimony, I reconstructed the evolutionary history of type IIL RE specificity. The number of specificity change events inferred from the tree in Figure 3 was 20 for the 1st base, 19 for the 2nd base, 19 for the 3rd base, 20 for the 4th base and 11 for the 6th base (Figure 4).

Next I reconstructed the ancestral states of the 7 amino acids that conferred restriction site specificity. In particular, I was interested in whether the amino acid substitutions that cause specificity changes in the mutagenesis experiment actually occurred during the history of evolution. Table 1 lists the specific amino acid substitution at each site (from the mutagenesis experiment), the number of substitution events of that particular substitution, and the total number of substitution events reconstructed using the phylogeny of 47 original sequences. My results showed that these substitutions not only happened during evolution but were the most prevalent substitution types at positions 645, 773, 806, 808 and 810. Remarkably, all the substitution events at positions 806 (K->E) and 808 (D->R) happened in perfect synchrony with specificity changes from G->C at base 6 (Figure 4). Moreover, the amino acid substitutions and specificity change happened multiples in parallel in the tree (Figure 4).

PAML analysis of the whole dataset

When the whole dataset was used (47 original sequences + 274 homologs), the likelihood ratio test indicated the existence of positive selection ($p < 0.01$, Table 2). However, the Bayesian approach did not identify any site with $\omega>1$. The overall distribution of ω is shown in Figure 5A. Relatively high ω values were mostly found within the methyltransferase domain and TRD, but no ω exceeded 1.

The Max dS in the whole dataset was 9.02 synonymous substitution per synonymous site (Table 2), which meant on average 9 synonymous substitutions per codon had occurred since the divergence of the sequence pair. It was quite possible that the sequences in the whole dataset were too divergent for PAML to work. To reduce the sequence divergence, I carried out PAML analyses in subclades of the tree.

Effect of sequence divergence

Among the 22 subclades of sequences that I tested, PAML detected positive selection in 20 subclades (likelihood ratio test, Table 2). However, only the subtree 2 consistently identified positively selected sites at different sequence divergence levels. The distributions of ω from the subtree datasets were similar to that of the whole dataset in that relative high ω appeared in the methyltransferase and TRD (Figure 5B).

The number of positive sites identified by PAML was sensitive to the sequence divergence. For subtree 2, PAML identified more sites as the sequences divergence decreased from max dS of 8.46 synonymous substitutions per synonymous site, peaking at max dS of 7.43 synonymous substitutions per synonymous site. Further decrease in the max sequence divergence resulted in fewer sites being identified, until finally no site was identified when max dS reached 1.91 synonymous substitutions per synonymous site. In general, sites identified at lower sequence divergence were subsets of sites identified at higher sequence divergence.

For the third subtree, although likelihood ratio tests were always significant along different sequence divergence, only the highest and the lowest divergent datasets identified sites. However, the sites were not consistent with each other (Table 2).

For 28 close sequences from *C. diphtheriae,* sequence divergence was substantially lower (max dS = 0.4 synonymous substitutions per synonymous site). I found no evidence of positive selection as there was no significant different likelihoods between the null and alternative models.

Consistency of PAML analysis

To explicitly test the consistency of PAML analysis, I carried out a bootstrap analysis with 100 pseudoreplicate datasets that were derived from 82 sequences in the subclade 2 of subtree 2. Each of the 100 pseudoreplicate datasets contained 50 sequences of similar sequence divergence (max dS, mean: 6.56 synonymous substitutions per synonymous site, standard deviation: 0.82 synonymous substitutions per synonymous site). PAML detected positive selection in each of the 100 pseudoreplicates (likelihood ratio test, $p<0.01$). The number of positively selected sites identified by PAML in each dataset ranged from 3 to 26. In total, 55 sites were identified and the average bootstrap support value was 20.5%. Among all the identified sites, 8 sites (594, 631, 656, 669, 708, 745, 767, 774) appeared in more than half of the 100 datasets (bootstrap > 50, Figure 6) and their locations in the structure of the enzyme were shown in Figure 7.

Results from the branch-site model

For the branch-site model, no matter which dataset (the 47 original sequences or the 10 original sequences, Table 2) was used, no positive selection was detected as there was no significant difference in the likelihood between the null and the alternative model.

Discussion

Evidence for positive selection in type IIL REs

My ancestral reconstruction of the restriction sites of type IIL REs showed there had been extensive changes of functions (i.e., specificity) during the history of evolution. Given the clear fitness benefit of changing specificity in response to evolving phages, these functional changes are most likely adaptive. Accordingly, using the likelihood ratio test, PAML consistently detected positive selection in the entire dataset and the subclades of sequences. However, PAML was less consistent in identifying sites under positive selection. Among 4 subtrees analyzed, PAML was able to identify sites only in two of them. Variation in sites detected by PAML among the 4 subtrees might be biological, i.e., in different bacterial lineages, different sites were under positive selection in response to different foreign DNAs in the environment. Alternatively, it suggests PAML is sensitive to the sampling bias in the dataset as discussed below.

Impact of sequence divergence

Sequence divergence had a large impact on positively selected sites identified by PAML. PAML identified no positively selected sites in datasets containing the most and least divergent sequences (i.e., the whole dataset and sequences from *C. diphtheriae*). Within subtree 2, zero and two sites were found in the least and most divergent datasets respectively, while more sites were identified in the medium divergent datasets. This was expected because neither too divergent sequences nor too similar sequences are informative. Simulation study by Yang's group has shown that PAML's accuracy and power were higher when analyzing medium and high divergent datasets (Anisimova et al., 2002).

It has also been shown that adding more sequences largely increased both the accuracy and power of PAML (Anisimova et al., 2002). As a result, larger datasets were more tolerant of higher sequence divergence than smaller datasets and there was no one cutoff for sequence divergence in PAML analyses. As seen in this study, the effect of sequence divergence in the more divergent datasets was mitigated by the increasing number of sequences, as the average dS increased at a lower rate than the max dS. For reference, I surveyed some of the previous studies in which positively selected sites had been identified. The sequence divergence used in those studies measured by max dS ranged from 0.15 to 4.45 synonymous substitutions per synonymous site, and measured by average dS ranged from 0.067 to 1.52 (Yang, 1998; Yang et al., 2000; Yang, Swanson & Vacquier, 2000; Yang & Swanson, 2002; Viljakainen & Pamilo, 2008; Areal et al., 2011). The divergence of the datasets in this study overlapped with previous studies but spanned a larger range (max dS ranged from 0.4 to 9.66, and average dS ranged from 0.21 to 2.79 synonymous substitutions per synonymous site). The more divergent datasets in my study were compensated by including many more (>50) sequences than the previous studies.

False positives and negatives

7 sites responsible for specificity changes at bases 2, 3, 4, 6 in type IIL restriction enzymes have been identified by mutagenesis experiments. Most of mutants at these sites retained comparable activities (Morgan & Luyten, 2009; Callahan et al., 2016), suggesting few if any sites other than these 7 amino acids contributed to sequence recognition. The 7 sites were all located within the TRD and close to DNA (Figure 1B). Moreover, extensive parallel changes at sites (645, 773, 806, 808, 810) strongly implicated the importance of the amino acid replacement in the functional adaptation. As

such, they represented the best candidates for positively selected sites in this enzyme and therefore were used as true positives to benchmark the performance of PAML.

Among the 17 sites identified by PAML, only one site (774) belonged to the set of true positive sites. All the other sites were either in the TRD or methyltransferase domain. None of them were located close to DNA in the structure or correlated significantly with specificity changes (Morgan & Luyten, 2009; Callahan et al., 2016). Because amino acids not making direct contact with DNA can also influence specificity (Lukacs et al., 2000), I cannot exclude the possibility that other positively selected sites exist. On the other hand, for the reason I discussed above, I think it is unlikely many more such sites exist, except for the unidentified sites responsible for specificity changes at the 1st base. One possible reason why only site 774 has been identified by PAML is that site 774 was under diversifying selection, while other 5 sites with functional support were mostly under directional selection (Table 1). Yang's group also proposed the possibility that PAML is better at detecting recurrent diversifying selection, but lack the power in directional selection detection (Anisimova et al., 2002). The little overlap between sites identified by PAML and the 7 true positive sites showed that PAML had clearly a high false negative rate and also quite possible a high false positive rate. The same problem has been reported in previous studies that tested PAML in systems with functional data (Yokoyama et al., 2008; Zhuang et al., 2009).

My results also suggest that neither the posterior probability nor bootstrap value is a good metric of confidence that can be placed on the sites identified by PAML. All the identified sites had a posterior probability of >0.95 to have $\omega>1$. Two of them (sites 708 and 656) also had very high bootstrap support values (> 80%).

Consistency of PAML

PAML was very consistent in detecting positive selection in type IIL REs. However, it was not so consistent in identifying positively selected sites. The average bootstrap value of the 55 identified sites was 20.5% and only 8 of them received a bootstrap support greater than 50%. The low bootstrap support suggested that signal used by PAML to identify positively selected sites was weak and PAML was very sensitive to the sampling bias in the dataset. Similarly, study using HA1 gene of human H3N2 influenza virus as the system also found similar inconsistency problem of PAML. They showed large variation among identified sites from seven highly overlapped datasets (Chen & Sun, 2011).

Conclusion

Using Type IIL as the model system, my results showed that PAML was very consistent at detecting positive selection in molecular sequences, but performed poorly in identifying positively selected sites. PAML failed to identify the vast majority of sites that were presumably under positive selection and instead identified many possible false positive sites, resulting in both low sensitivity and specificity. Given the likely high false positive rate, we should treat sites identified by PAML with caution and validate them with functional data whenever possible.

References

Aguileta, Gabriela, Guislaine Refrégier, Roxana Yockteng, Elisabeth Fournier, and Tatiana Giraud, ‘Rapidly Evolving Genes in Pathogens: Methods for Detecting Positive Selection and Examples among Fungi, Bacteria, Viruses and Protists’, *Infection, Genetics and Evolution*, 9 (2009), 656–70

Akey, Joshua M., ‘Constructing Genomic Maps of Positive Selection in Humans: Where Do We Go from Here?’, *Genome Research*, 2009, 711–22

Akey, Joshua M., Ge Zhang, Kun Zhang, Li Jin, and Mark D. Shriver, ‘Interrogating a High-Density SNP Map for Signitures of Natural Selection’, *Genome Research*, 2002, 1805–14

Anisimova, M., J. P. Bielawski, and Z. Yang, ‘Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution’, *Molecular Biology and Evolution*, 18 (2001), 1585–92

Anisimova, Maria, Joseph P Bielawski, and Ziheng Yang, ‘Accuracy and Power of Bayes Prediction of Amino Acid Sites under Positive Selection’, *Molecular Biology and Evolution*, 19 (2002), 950–58

Areal, Helena, Joana Abrantes, and Pedro J Esteves, ‘Signatures of Positive Selection in Toll-like Receptor (TLR) Genes in Mammals’, *BMC Evolutionary Biology*, 11 (2011), 368

Barreiro, Luis B., Guillaume Laval, Hélène Quach, Etienne Patin, and Lluís Quintana-Murci, ‘Natural Selection Has Driven Population Differentiation in Modern Humans’, *Nature Genetics*, 40 (2008), 340–45

Begun, David J., Alisha K. Holloway, Kristian Stevens, La Deana W. Hillier, Yu Ping Poh, Matthew W. Hahn, and others, ‘Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila Simulans’, *PLoS Biology*, 5 (2007), 2534–59

Bierne, Nicolas, and Adam Eyre-Walker, ‘The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila’, *Molecular Biology and Evolution*, 21 (2004), 1350–60

Callahan, Scott J., Yvette A. Luyten, Yogesh K. Gupta, Geoffrey G. Wilson, Richard J. Roberts, Richard D. Morgan, and others, ‘Structure of Type IIL Restriction-Modification Enzyme MmeI in Complex with DNA Has Implications for Engineering New Specificities’, *PLoS Biology*, 14 (2016), 1–18

Charlesworth, J, and A Eyre-Walker, ‘The Rate of Adaptive Evolution in Enteric Bacteria’, *Molecular Biology and Evolution*, 23 (2006), 1348–56

Chen, Jiming, and Yingxue Sun, ‘Variation in the Analysis of Positively Selected Sites Using Nonsynonymous/synonymous Rate Ratios: An Example Using Influenza Virus’, *PLoS ONE*, 6 (2011)

Doniger, Scott W., Hyun Seok Kim, Devjanee Swain, Daniella Corcuera, Morgan Williams, Shiaw Pyng Yang, and others, ‘A Catalog of Neutral and Deleterious Polymorphism in Yeast’, *PLoS Genetics*, 4 (2008)

Eyre-Walker, Adam, ‘Changing Effective Population Size and the McDonald-Kreitman Test’, *Genetics*, 162 (2002), 2017–24

Ford, Michael J., ‘Applications of Selective Neutrality Tests to Molecular Ecology’, *Molecular Ecology*, 11 (2002), 1245–62

Galtier, Nicolas, ‘Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis’, *PLoS Genetics*, 12 (2016)

Goldman, Nick, and Ziheng Yangf-, ‘A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences.’, *Molecular Biology and Evolution*, 1994

Hahn, Matthew W., ‘Toward a Selection Theory of Molecular Evolution’, *Evolution*, 2008, 255–65

Heger, Andreas, and Chris P. Ponting, ‘Evolutionary Rate Analyses of Orthologs and Paralogs from 12 Drosophila Genomes’, *Genome Research*, 17 (2007), 1837–49

Hohenlohe, Paul A., Susan Bassham, Paul D. Etter, Nicholas Stiffler, Eric A. Johnson, and William A. Cresko, ‘Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags’, *PLoS Genetics*, 6 (2010)

Hughes, Austin L., and Robert Friedman, ‘Codon-Based Tests of Positive Selection, Branch Lengths, and the Evolution of Mammalian Immune System Genes’, *Immunogenetics*, 60 (2008), 495–506

Hughes, Austin L., and Meredith Yeager, ‘Natural Selection At Major Histocompatibility Complex Loci of Vertebrates’, *Annual Review of Genetics*, 32 (1998), 415–35

Katoh, Kazutaka, and Daron M. Standley, ‘MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability’, *Molecular Biology and Evolution*, 30 (2013), 772–80

Kimura, Motoo, *The Neutral Theory of Molecular Evolution*, 1983

Lewontin, R C, and Jesse Krakauer, ‘Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms’, *Genetics*, 74 (1973), 175–95

Maddison, W P, and D R Maddison, ‘Mesquite: A Modular System for Evolutionary Analysis. Version Version 3.31.’, *URL Http://mesquiteproject. Org*, 2016

McDonald, John H., and Martin Kreitman, ‘Adaptive Protein Evolution at the Adh Locus in Drosophila’, *Nature*, 351 (1991), 652–54

Metzger, Kelsey J, and Michael a Thomas, ‘Evidence of Positive Selection at Codon Sites Localized in Extracellular Domains of Mammalian CC Motif Chemokine Receptor Proteins.’, *BMC Evolutionary Biology*, 10 (2010), 139

Morgan, Richard D., and Yvette A. Luyten, ‘Rational Engineering of Type II Restriction Endonuclease DNA Binding and Cleavage Specificity’, *Nucleic Acids Research*, 37 (2009), 5222–33

Murray, Noreen E., Anne S. Daniel, Gill M. Cowan, and Paul M. Sharp, ‘Conservation of Motifs within the Unusually Variable Polypeptide Sequences of Type I Restriction and Modification Enzymes’, *Molecular Microbiology*, 9 (1993), 133–43

Myles, S., K. Tang, M. Somel, R. E. Green, J. Kelso, and M. Stoneking, ‘Identification and Analysis of Genomic Regions with Large between-Population Differentiation in Humans’, *Annals of Human Genetics*, 72 (2008), 99–110

Namroud, Marie Claire, Jean Beaulieu, Nicolas Juge, Jérôme Laroche, and Jean Bousquet, ‘Scanning the Genome for Gene Single Nucleotide Polymorphisms Involved in Adaptive Population Differentiation in White Spruce’, *Molecular Ecology*, 17 (2008), 3599–3613

Nielsen, Einar E, Jakob Hemmer-Hansen, Nina A Poulsen, Volker Loeschcke, Thomas Moen, Torild Johansen, and others, ‘Genomic Signatures of Local Directional Selection in a High Gene Flow Marine Organism; the Atlantic Cod (Gadus Morhua)’, *BMC Evolutionary Biology*, 9 (2009), 276

Park, Seung Gu, Donghyun Park, Yu Jin Jung, Eunkyung Chung, and Sun Shim Choi, ‘Positive Selection Signatures in the TLR7 Family’, *Genes and Genomics*, 32 (2010), 143–50

Roberts, Richard J., Tamas Vincze, Janos Posfai, and Dana Macelis, ‘REBASE-a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes’, *Nucleic Acids Research*, 43 (2015), D298–99

Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, and others, ‘Detecting Recent Positive Selection in the Human Genome from Haplotype Structure’, *Nature*, 419 (2002), 832–37

Sabeti, PC, Patrick Varilly, Ben Fry, and Jason Lohmueller, ‘Genome-Wide Detection and Characterization of Positive Selection in Human Populations’, *Nature,* 449 (2007), 913–18

Schlüns, Helge, and Ross H. Crozier, ‘Molecular and Chemical Immune Defenses in Ants (Hymenoptera: Formicidae)’, *Myrmecological News,* 12 (2009), 237–49

Sharp, P M, J E Kelleher, A S Daniel, G M Cowan, and N E Murray, ‘Roles of Selection and Recombination in the Evolution of Type I Restriction-Modification Systems in Enterobacteria’, *Proc Natl Acad Sci USA,* 89 (1992), 9836–40

Smith, Nick G C, and Adam Eyre-Walker, ‘Adaptive Protein Evolution in Drosophila’, *Nature,* 415 (2002), 1022–24

Stamatakis, Alexandros, ‘RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies’, *Bioinformatics,* 30 (2014), 1312–13

Stern, Adi, and Rotem Sorek, ‘The Phage-Host Arms Race: Shaping the Evolution of Microbes’, *BioEssays,* 33 (2011), 43–51

Strasburg, Jared L., Caroline Scotti-Saintagne, Ivan Scotti, Zhao Lai, and Loren H. Rieseberg, ‘Genomic Patterns of Adaptive Divergence between Chromosomally Differentiated Sunflower Species’, *Molecular Biology and Evolution,* 26 (2009), 1341–55

Sundaram, Arvind Y M, Sonia Consuegra, Viswanath Kiron, and Jorge M O Fernandes, ‘Positive Selection Pressure within Teleost Toll-like Receptors tlr21 and tlr22 Subfamilies and Their Response to Temperature Stress and Microbial Components in Zebrafish’, *Molecular Biology Reports,* 39 (2012), 8965–75

Suzuki, Yoshiyuki, and Masatoshi Nei, ‘Simulation Study of the Reliability and Robustness of the Statistical Methods for Detecting Positive Selection at Single Amino Acid Sites.’, *Molecular Biology and Evolution,* 19 (2002), 1865–69

Tschirren, B., L. Råberg, and H. Westerdahl, ‘Signatures of Selection Acting on the Innate Immunity Gene Toll-like Receptor 2 (TLR2) during the Evolutionary History of Rodents’, *Journal of Evolutionary Biology,* 24 (2011), 1232–40

Vasu, K., and V. Nagaraja, ‘Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense’, *Microbiology and Molecular Biology Reviews,* 77 (2013), 53–72

Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard, ‘A Map of Recent Positive Selection in the Human Genome’, *PLoS Biology,* 4 (2006), 0446–58

Williamson, Hubisz, Andrew G Clark, Payseur, Carlos D Bustamante, and Nielsen, ‘Localizing Recent Adaptive Evolution in the Human Genome’, *PLoS Genetics*, 3 (2007), e90

Wong, Wendy S W, Ziheng Yang, Nick Goldman, and Rasmus Nielsen, ‘Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites’, *Genetics*, 168 (2004), 1041–51

Wu, Martin, Sourav Chatterji, and Jonathan A. Eisen, ‘Accounting for Alignment Uncertainty in Phylogenomics’, *PLoS ONE*, 7 (2012)

Yang, Z, ‘Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A’, *Journal of Molecular Evolution*, 51 (2000), 423–32

Yang, Z, ‘Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution’, *Mol Biol Evol*, 15 (1998), 568–73

Yang, Z., and W. J. Swanson, ‘Codon-Substitution Models to Detect Adaptive Evolution That Account for Heterogeneous Selective Pressures Among Site Classes’, *Molecular Biology and Evolution*, 19 (2002), 49–57

Yang, Z., W. J. Swanson, and V. D. Vacquier, ‘Maximum-Likelihood Analysis of Molecular Adaptation in Abalone Sperm Lysin Reveals Variable Selective Pressures Among Lineages and Sites’, *Molecular Biology and Evolution*, 17 (2000), 1446–55

Yang, Ziheng, ‘PAML 4: Phylogenetic Analysis by Maximum Likelihood’, *Molecular Biology and Evolution*, 24 (2007), 1586–91

Yang, Ziheng, and Joseph R. Bielawski, ‘Statistical Methods for Detecting Molecular Adaptation’, *Trends in Ecology and Evolution*, 2000, 496–503

Yang, Ziheng, and Mario Dos Reis, ‘Statistical Properties of the Branch-Site Test of Positive Selection’, *Molecular Biology and Evolution*, 28 (2011), 1217–28

Yokoyama, S., T. Tada, H. Zhang, and L. Britt, ‘Elucidation of Phenotypic Adaptations: Molecular Analyses of Dim-Light Vision Proteins in Vertebrates’, *Proceedings of the National Academy of Sciences*, 105 (2008), 13480–85

Zhang, Liqing, and Wen Hsiung Li, ‘Human SNPs Reveal No Evidence of Frequent Positive Selection’, *Molecular Biology and Evolution*, 22 (2005), 2504–7

Zheng, Yu, Richard J. Roberts, and Simon Kasif, ‘Identification of Genes with Fast-Evolving Regions in Microbial Genomes’, *Nucleic Acids Research*, 32 (2004), 6347–57

Zheng, Yu, Richard J. Roberts, and Simon Kasif, ‘Identification of Genes with Fast-Evolving Regions in Microbial Genomes’, *Nucleic Acids Research*, 32 (2004), 6347–57

Zhuang, Hanyi, Ming-Shan Chien, and Hiroaki Matsunami, ‘Dynamic Functional Evolution of an Odorant Receptor for Sex-Steroid-Derived Odors in Primates.’, Proceedings of the National Academy of Sciences of the United States of America, 106 (2009), 21247–51

Tables

Table 1. Evolutionary changes of the 7 amino acid sites conferring specificity.

Recognition site base	Amino acid	Amino acid substitution	# of change/# of total change
2nd	645	K <--> M	11/21
3rd	751	E <--> R	1/28
	773	N <--> D	15/19
4th	774	A <--> K	0/31
	810	R <--> S	7/12
6th	806	E <--> K	10/12
	808	R <--> D	10/11

Table 2. Evidence for positive selection and positively selected sites identified by PAML

Dataset		Total # (original sequence # + homolog #)	2ΔlnL M7 vs. M8	Sites by M8	Max dS (synonymous substitution per synonymous site)	Average dS (synonymous substitution per synonymous site)
Whole dataset:		321 (47+274)	50.70**	None	9.02	2.76
Sequence divergence analysis						
Subtree 1	Subclade 1	90 (5+85)	0	Not applicable	9.66	2.51
	Subclade 2	69 (4+65)	28.60**	None	7.45	2.51
	Subclade 3	55 (4+51)	245.84**	None	6.99	2.43
	Subclade 4	42 (2+40)	550.88**	None	5.81	2.58
	Subclade 5	19 (2+17)	173.31**	None	4.64	2.06
	Subclade 6	13 (1+12)	338.59**	None	4.21	2.04
Subtree 2	Subclade 1	186 (10+176)	303.97**	656, 745	8.46	2.47
	Subclade 2	82 (7+75)	114.18**	594, 631, 656, 666, 669, 706, 708, 745, 767, 774	7.43	2.37
	Subclade 3	57 (6+51)	81.59**	594, 630, 656, 669, 708, 745, 774	6.79	2.34
	Subclade 4	47 (4+43)	90.85**	594, 630, 631, 632, 669, 709, 745, 774	6.79	2.05

	Subclade 5	29 (3+26)	116.56**	630, 632, 709, 745	4.69	1.51
	Subclade 6	22 (1+21)	154.75**	557, 630, 632, 745	2.17	1.22
	Subclade 7	14 (1+13)	2371.86**	None	1.91	1.20
Subtree 3	Subclade 1	71 (8+63)	447.39**	713	7.76	2.76
	Subclade 2	46 (6+40)	409.99**	None	8.73	2.51
	Subclade 3	32 (2+30)	581.24**	None	6.91	2.38
	Subclade 4	20 (1+19)	38.06**	None	5.34	2.26
	Subclade 5	9 (1+8)	478.69**	591, 635	4.76	2.62
Subtree 4	Subclade 1	55 (4+51)	25.39**	None	6.04	2.60
	Subclade 2	48 (3+45)	0	Not applicable	6.87	2.66
	Subclade 3	31 (2+29)	32.52**	None	5.41	2.66
	Subclade 4	12 (2+10)	290.46**	None	4.44	2.66

Sequences from *C. diphtheriae*	30 (2+28)	0.53	Not applicable	0.40	0.21
Branch-site model analysis					
Whole dataset	47	0	Not applicable	6.90	2.79
Subtree dataset	10	0	Not applicable	3.00	2.58

2ΔlnL (twice the log likelihood difference between models) is compared to a χ2 distribution with 2 degrees of freedom (critical values 5.99, 9.21 at 5% and 1% significance respectively, '**' means p<0.01). Only sites with probability higher than 0.95 to be with ω > 1 are included (sites with probability higher than 0.99 are underlined).

Figure Legend

Figure 1. The structure of MmeI enzyme-DNA complex. A) side view with the 6 bases of the recognition site labeled. B) top view. 7 amino acids that recognize the 6 bases are labeled in black. The endonuclease domain of the enzyme is not shown.

Figure 2. Phylogeny of Type IIL RE 796 sequences (47 original sequence + 749 homologs from genomes of 189 bacterial species). The original sequences are labeled along with the four subtrees used in the sequence divergence analysis.

Figure 3. Rapid evolution of recognition sites in type IIL restriction enzymes. The left is the phylogeny of 47 original sequences. The right is the alignment of recognition sites, with different nucleotides highlighted in different colors. The number of evolutionary changes in every base of recognition site is summarized in the first row. 10 original sequences used for branch-site model analysis are in red.

Figure 4. Evolutionary changes of amino acid sites (806, 808) and 6th base of the recognition site. Change (G->C) at 6th base and amino acid changes at 806 (K->E), 808 (D->R) were concurrent and were labeled in green on the phylogeny of the 47 original sequences. Other type of amino acid substitutions were labeled in red. No changes were inferred on unlabeled branches.

Figure 5. Overall distribution of ω among all sites in the sequence alignment with five domains differently colored. A) the whole dataset containing the 47 original sequences and 274 homologs. B) Subclade 2 of subtree 2 (7 original sequences and 75 homologs). 10 sites identified to be under positive selection are circled in red.

Figure 6. Result of PAML bootstrap analysis with 100 pseudoreplicates. Amino acid positions (sites) of MmeI enzyme are shown on the X-axis. Y-axis shows the number of times a site was identified under positive selection by PAML. Sites with greater than 50% bootstrap support are labeled in red.

Figure 7. Positions of the 8 most frequently identified positively selected sites (labeled in red) from the bootstrap analysis.

Figure 1.

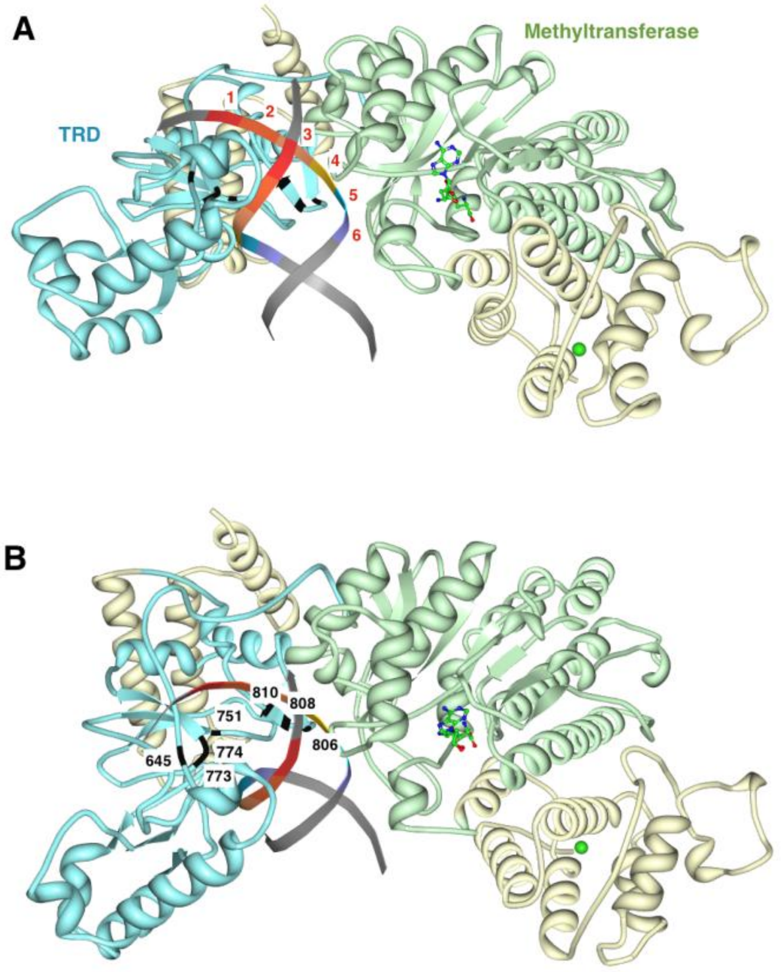


Figure 2.

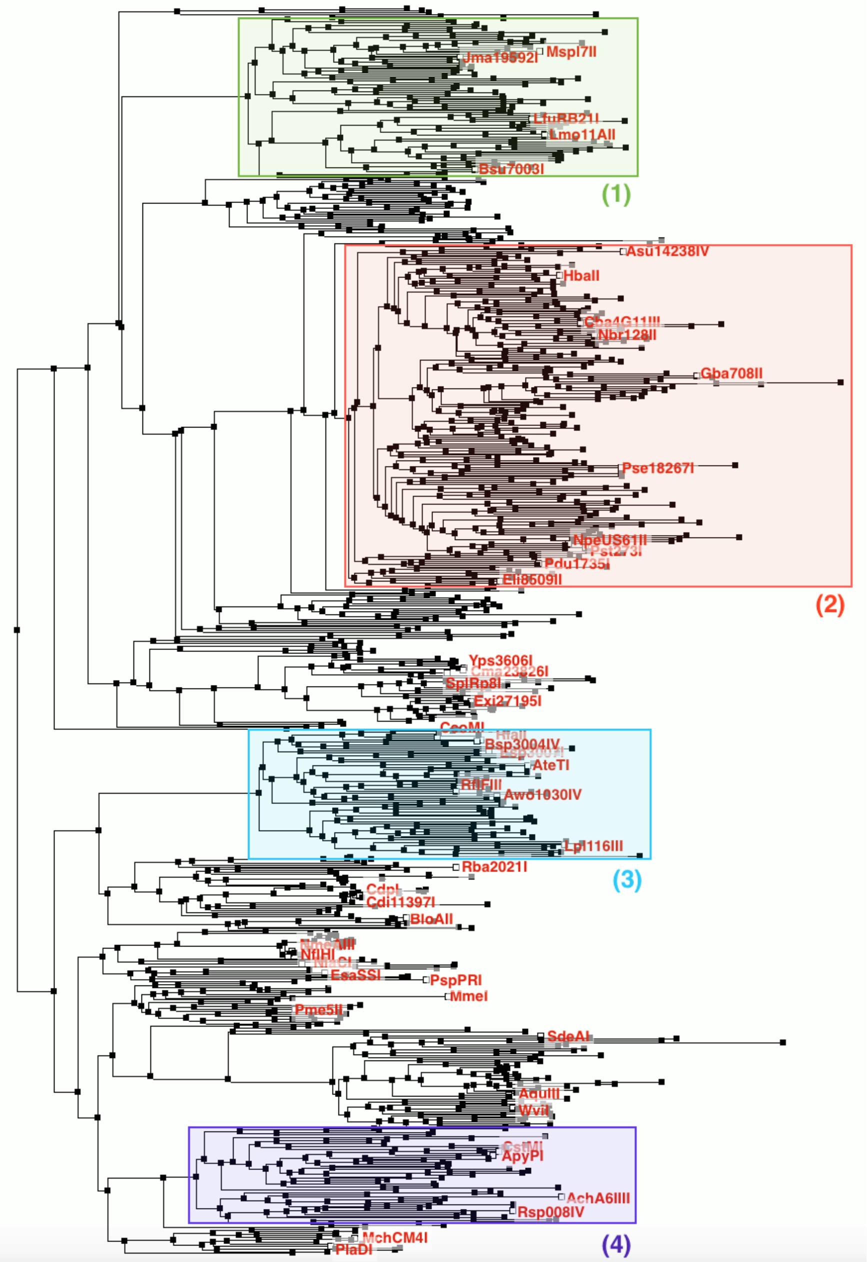
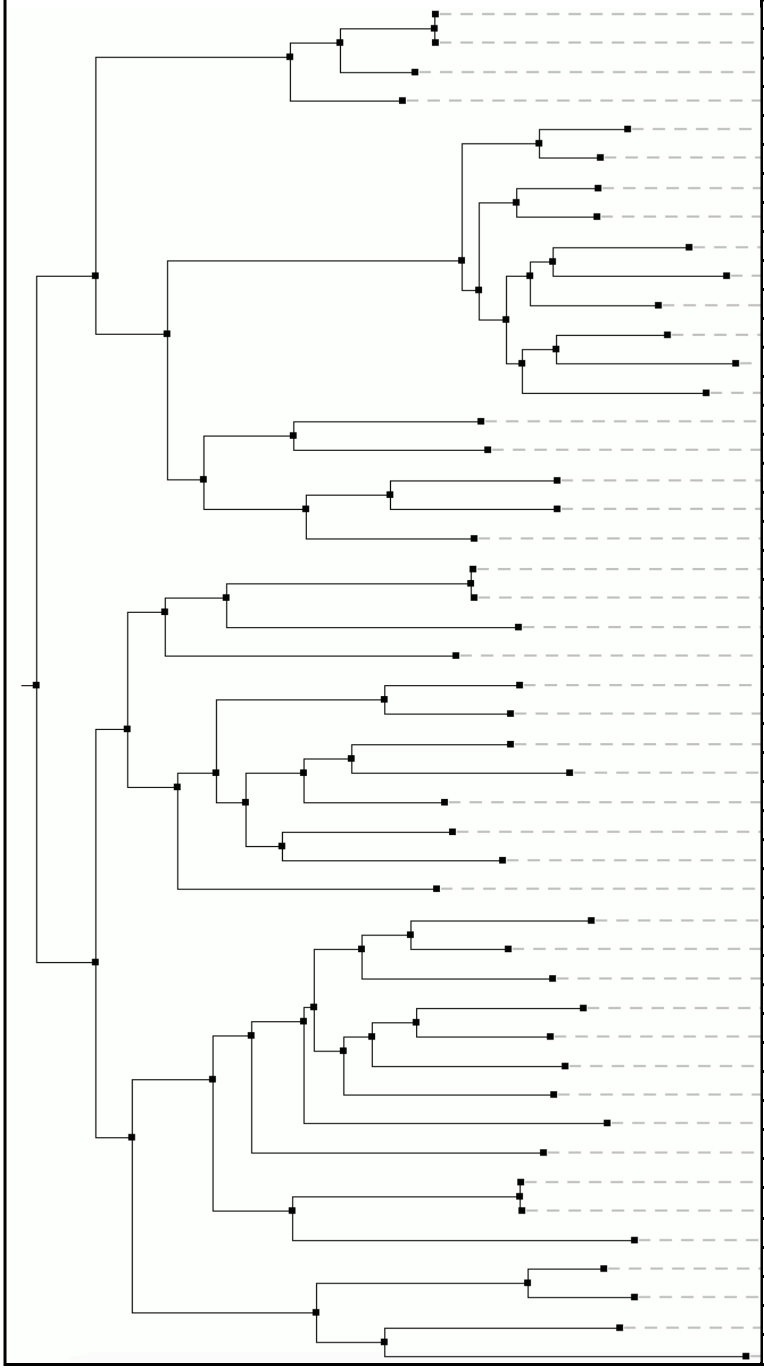

MspI7II
Jma19592I
LfuRB21I
Lmo11AII
Bsu7003I
(1)
Asu14238IV
HbaII
Cba4G11III
Nbr128II
Gba708II
Pse18267I
NpeUS61II
Pst273I
Pdu1735I
Eli8509II
(2)
Yps3606I
Cma23826I
SplRp8I
Exi27195I
CcoMI
HfaII
Bsp3004IV
Esp3007I
AteTI
RflFIII
Awo1030IV
Lpl116III
(3)
Rba2021I
CdpI
Cdi11397I
BloAII
NmeAIII
NflHI
NlaCI
EsaSSI
PspPRI
MmeI
Pme5II
SdeAI
AquIII
WviI
CstMI
ApyPI
AchA6III
Rsp008IV
(4)
MchCM4I
PlaDI

Figure 3.



		1	2	3	4	5	6
# of changes		20	19	19	20	0	11
Cma23826I		C	G	G	A	A	G
Yps3606I		C	G	G	A	A	G
SplRp8I		C	C	C	A	A	G
Exi27195I		G	C	C	G	A	C
NpeUS61II	G	A	T	C	G	A	C
Pst273I	G	A	T	C	G	A	G
Cba4G11III		C	A	C	G	A	G
Nbr128II		A	C	C	G	A	C
Eli8509II		C	C	G	G	A	G
Pse18267I	R	C	C	G	A	A	G
HbaII		G	C	C	C	A	G
Pdu1735I		C	A	C	C	A	C
Gba708II		A	T	G	C	A	C
Asu14238IV		C	G	T	R	A	C
Jma19592I	G	T	A	T	N	A	C
MspI7II		A	C	G	R	A	G
LfuRB21I		A	A	G	G	A	G
Lmo11AII		T	A	G	R	A	G
Bsu7003I		G	A	C	G	A	G
NmeAIII		G	C	C	G	A	G
NflHI		G	C	G	G	A	G
NlaCI		C	A	T	C	A	C
PspPRI		C	C	Y	C	A	G
PlaDI		C	A	T	C	A	G
MchCM4I		G	A	G	G	A	G
WviI		C	A	C	R	A	G
SdeAI		C	A	G	R	A	G
AquIII		G	A	G	G	A	G
Pme5II		G	A	C	G	A	G
MmeI		T	C	C	R	A	C
EsaSSI		G	A	C	C	A	C
RlaII		A	C	A	C	A	G
CcoMI		C	A	G	C	A	G
Bsp3004IV		C	C	G	C	A	T
Awo1030IV		G	C	C	R	A	G
RflFIII		C	G	C	C	A	G
AteTI		G	G	G	R	A	G
Esp3007I		C	A	G	A	A	G
Lpl116III		C	A	G	R	A	G
Rba2021I		C	A	C	G	A	G
CdpI		G	C	G	G	A	G
Cdi11397I		G	C	G	C	A	G
BloAII		G	A	G	G	A	C
CstMI		A	A	G	G	A	G
ApyPI		A	T	C	G	A	C
Rsp008IV		A	C	G	C	A	G
AchA6III		A	G	C	C	A	G

Figure 4.

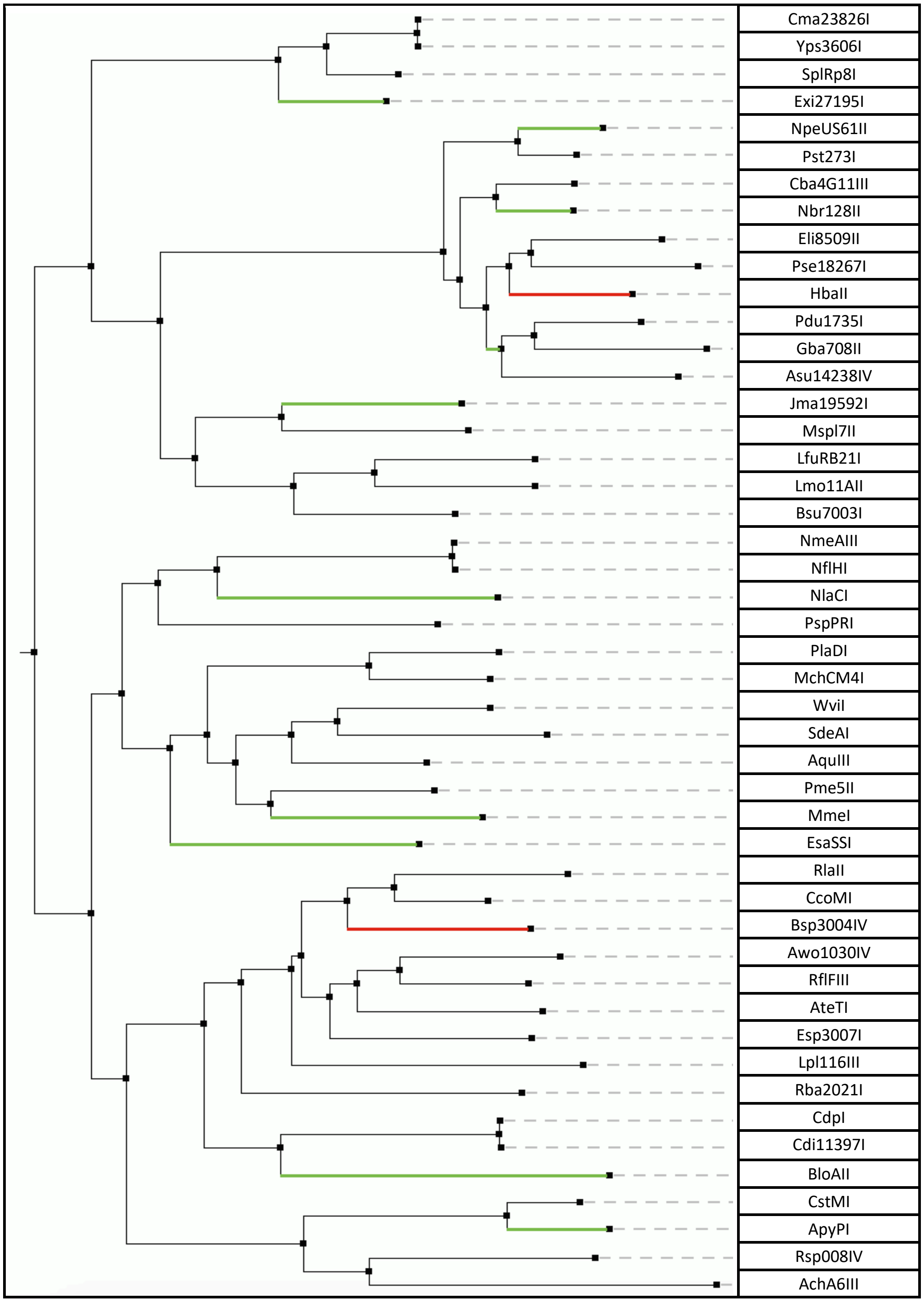

Cma23826I
Yps3606I
SplRp8I
Exi27195I
NpeUS61II
Pst273I
Cba4G11III
Nbr128II
Eli8509II
Pse18267I
HbaII
Pdu1735I
Gba708II
Asu14238IV
Jma19592I
MspI7II
LfuRB21I
Lmo11AII
Bsu7003I
NmeAIII
NflHI
NlaCI
PspPRI
PlaDI
MchCM4I
WviI
SdeAI
AquIII
Pme5II
MmeI
EsaSSI
RlaII
CcoMI
Bsp3004IV
Awo1030IV
RflFIII
AteTI
Esp3007I
Lpl116III
Rba2021I
CdpI
Cdi11397I
BloAII
CstMI
ApyPI
Rsp008IV
AchA6III

Figure 5.

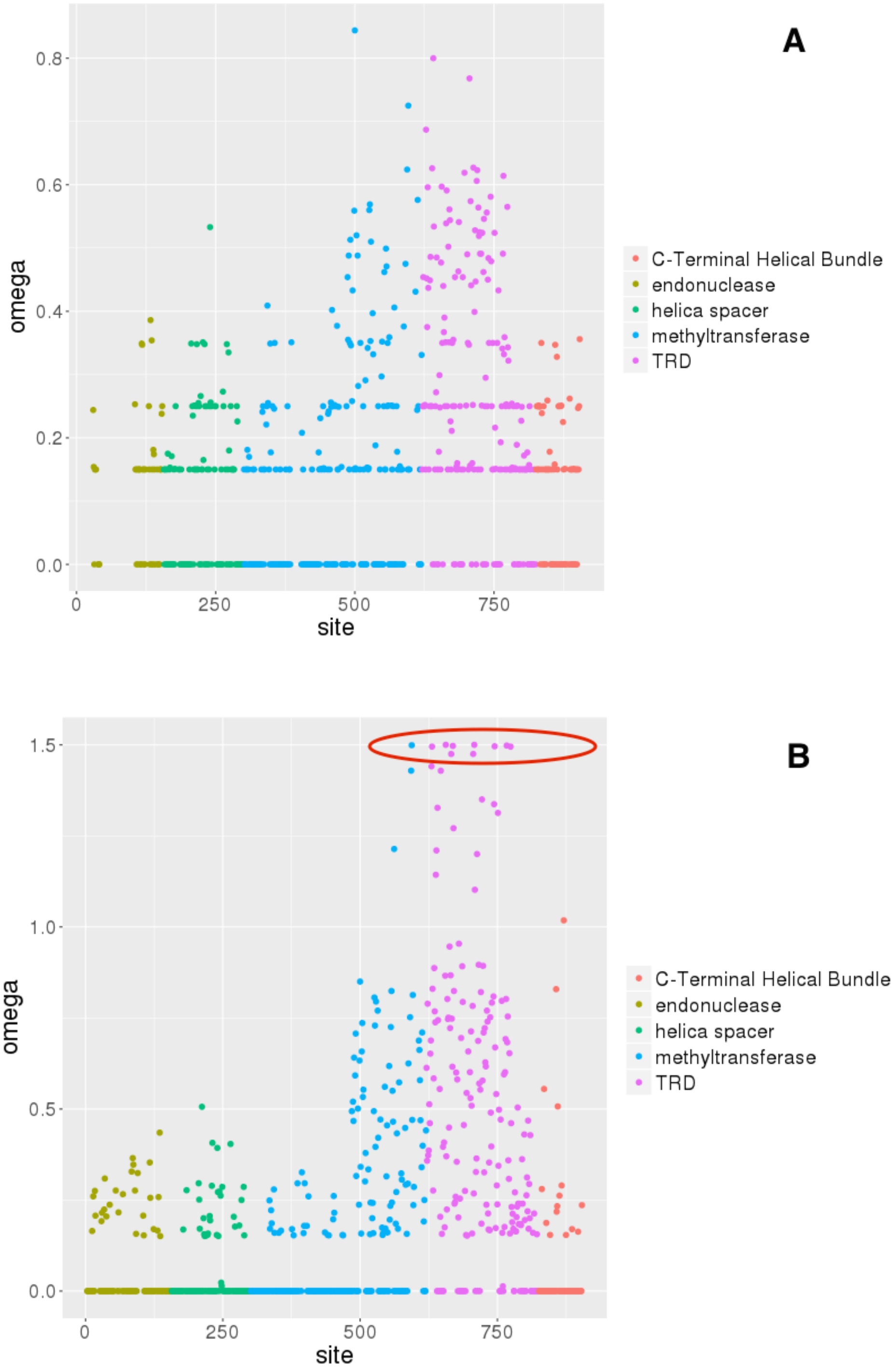

A
0.8
0.6
0.4
0.2
0.0
omega
0
250
500
750
site
● C-Terminal Helical Bundle
● endonuclease
● helica spacer
● methyltransferase
● TRD
B
1.5
1.0
0.5
0.0
omega
0
250
500
750
site
● C-Terminal Helical Bundle
● endonuclease
● helica spacer
● methyltransferase
● TRD

Figure 6.

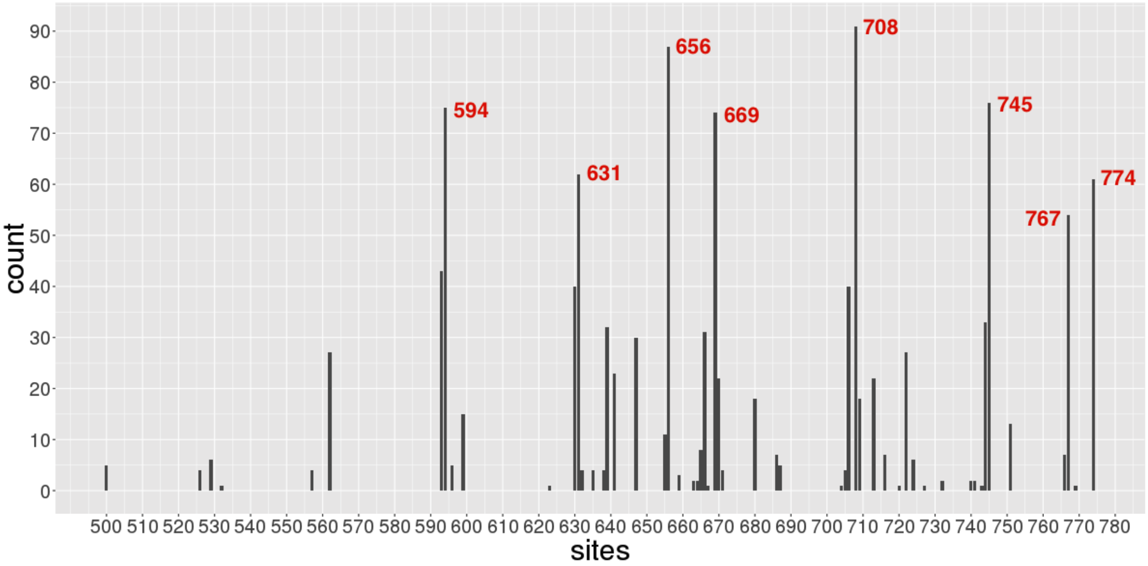
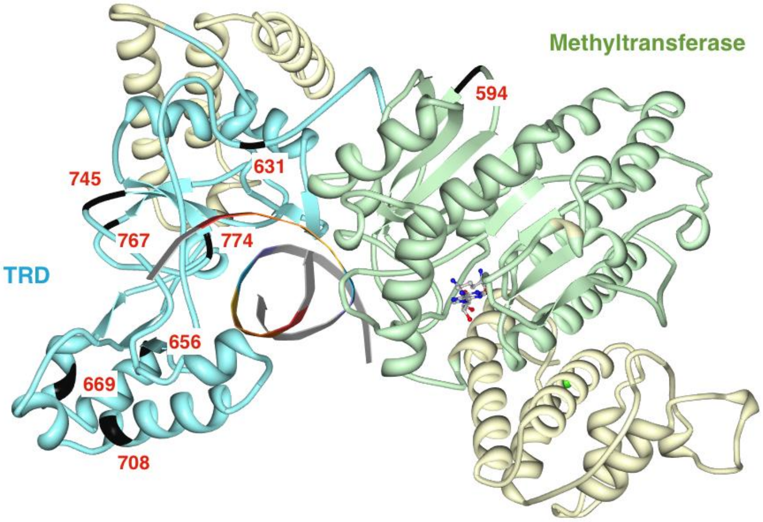

count
90
80
70
60
50
40
30
20
10
0
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
660
670
680
690
700
710
720
730
740
750
760
770
780
sites
594
631
656
669
708
745
767
774

Figure 7.


Methyltransferase
594
745
631
767
774
TRD
656
669
708