# PHYLOGENETIC CLUSTERING OF DNA POL IV/V PROTEINS AND DETECTION OF THE LITTLE FINGER DOMAIN FOR IMPROVED CLASSIFICATION OF Y-FAMILY POLYMERASES

LOVING, Joshua[1], TASSINARI, Anna[1], HERNANDEZ, Yozen[1], and LOECHLER, Edward [1,2],
(1) Graduate Program in Bioinformatics (2) Department of Biology, Boston University, 24 Cummington Street, Boston, MA 02215

BOSTON UNIVERSITY

**Abstract:** Cells suffer damage to their DNA from a variety of external factors including ionizing radiation and DNA-binding chemicals. Some types of DNA damage can cause replication to halt, as the DNA polymerase is unable to bypass the physical change to the DNA molecule. In order to cope with certain types of genetic damage, living systems have evolved damage-induced DNA polymerases capable of identifying and continuing translation at these lesion sites through translesion synthesis (TLS). The DNA Polymerase (DNA Pol) Y-Family proteins are all damage-inducible polymerases which share little similarity with other polymerases but closely resemble those within the family. Members of the human DNA Pol η class, a protein in the Y-Family, typically respond to UV DNA damage. When recruited to repair damage caused by an intercalating agent, such as Benzo[a]pyrene, a Pol η may occasionally erroneously insert an incorrect nucleotide, resulting in a mutation which may lead to a cancerous cell. DNA Pol κ, another Y-Family class, on the other hand successfully repairs the damage without causing a mutation. By examining key features of the bacterial orthologs of these proteins, in particular the binding-assisting "little finger" domain, we may be able to isolate significant differences in sequence which may explain how these proteins interact with damaged DNA. While crystal structures for Pol IV (κ homolog) and κ exist, none have been successfully produced for Pol V (η homolog). By applying dynamic programming techniques and phylogenetic analysis we developed algorithms to address two problems: identification of the "little finger" domain of Pol IV and imputation of Pol V structure given knowledge of Pol IV, respectively.

## INTRODUCTION

**Y-family DNA Polymerases wish well but are not perfect.** Cells exposed to DNA-damaging factors cannot successfully undergo mitosis or transcription. In case a lesion is detected, a specialized **Y-family DNA Polymerase** is recruited to continue transcription past the lesion site. These polymerases are error prone and often mismatch the complementary to G amino acid causing mutation (**Fig. 1.**)

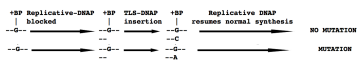


**Figure 1**. Incorrect insertion of A in stead of C during adduct bypass causes mutation

**Little Finger Domain (LF) (Fig. 2.) is crucial in bypassing lesions.** It increases protein-DNA interaction, stabilizing complex with the polymerases' catalytic region (fingers, palm, and thumb) allowing it to bind to "bulges" in DNA.

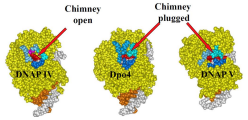


**Figure 3.** Predicted secondary structure showing adducts clogging the chimney in Dpo4 and DNAP V

**To successfully bypass adducts one needs:**

1. Large opening (**unplugged chimney**) (**Fig. 3.**)
2. **Second non-covalent bridge** to anchor Little Finger (**Fig. 4.A**) when bridge near active site is broken to allow bypass of bulky adducts which protrude into the minor grove (**Fig. 4.B**)

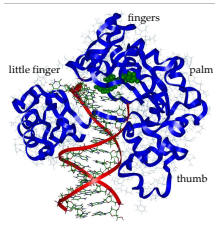


**Figure 2.** DNA Polymerase 3-D structure

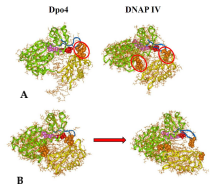


**Figure 4.** **(A)** Non-covalent bridges anchoring the Little Finger shown within predicted secondary structure of DNA Polymerases. **(B)** Breaking the active-site bridge in DNAP IV to allow larger adducts into the minor grove

**What we were hoping to accomplish:**

1. Figure out from sequence analysis diversity in Pol IV homologs. If Pol IV homologs can be further subdivided, elements unique to a subclass can be examined for selection, or for true membership in Pol IV.
2. Figure out similarities in Pol IV and Pol V sequences that can be used to infer a structure for Pol V, for which no crystal structure has yet been produced.

## METHODS

Data collection: Starting set of sequence identifiers (UniProt IDs) provided by PI as part of earlier study. Sequence data collected from the Protein database on NCBI's site. Total collection of annotated Pol IV sequences retrieved from the SuperFAM database.

| Dataset | # of sequences | Mapped NCBI GIs | # of duplicates |
|---|---|---|---|
| Starting set | 447 | 448 | 1 |
| All annotated as Pol IV | 2596 | 4779 | 2107 |

**Alignments:** Peptide sequences from starting dataset were aligned using CLUSTAL Omega (or traditional CLUSTAL for nucleotide alignments). Later, MUSCLE and T-Coffee were also used to include structural information in alignments.

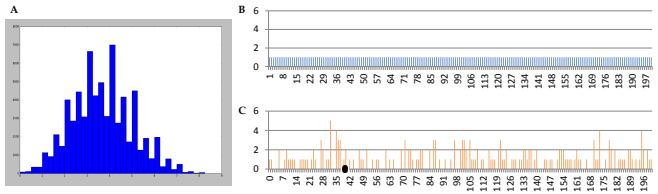


**Figure 5.** **(A)** Histogram of background distribution of the scores **(B)** Example initial positioning of sequences **(C)** Random reselection causes differential frequency distribution

**Little Finger Finder:** As previously described, the little finger regions of PolIV and PolV are important to their function because the structure of the little finger helps determine which nucleotide will be inserted. We exploited this fact by creating **a model of the amino acid properties needed in each alpha sheet and beta helix** necessary to produce the required structural functionality. Using this model, we created a scoring algorithm to generate a score for each candidate subregion. The distribution of scores for each candidate across all input sequences was used to select filter for regions with low p-value (Fig 5A). After using a Monte Carlo method to generate multiple input sets (Fig 5B,C), we used a dynamic programming algorithm to **find little finger regions with highest score based on maximizing subregion scores and minimizing turn lengths**.

**Phylogenetic Classification:** Sequences in our dataset were previously clustered based on amino acids on positions 40 and 41 of the alignment, which are crucial to Pol IV structure and function. **We started by building eight phylogenetic trees using distinct methods** (**Table 1.B and C**) and both amino-acid as well as nucleotide sequences. To check if an overlap of clades exists with the pre-determined grouping, we constructed a consensus tree replacing leaf labels with group assignment (**Table 1.C**). Next, **we added several additional positions of interest** to see if the classification of Pol IV and Pol V could be further resolved (**Table 1.D**). Amino acids at these new locations correspond to structural characteristics crucial for Pol IV and V function:

1. **G31** is important because the presence of **anything except a G** at that position would produce a **"plugged" chimney**, which is not characteristic of a Pol IV member.
2. **Amino acids larger than G or P on position 74** may also **obstruct the chimney** in polymerases.
3. **Position 250** is important because as long as it is occupied by **anything else than K, R or Q**, the **bridge between the little finger and the rest of the protein can easily be broken** – a distinctive feature of Pol IV.

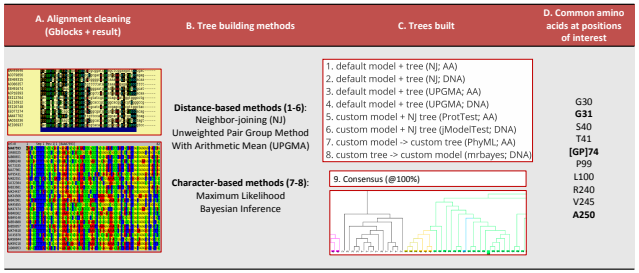


**Table 1.** Outline of phylogenetic analysis steps and methods used for building trees

## RESULTS

1. The early attempts to resolve differentiation of Pol IV and Pol V sequences based on the G31 position were unsuccessful. We found multiple sequences with a "contradictory" pair of features (i.e., G and K, or not G and not K/R/Q) mixed in among either Pol Iv or Pol V, making it more difficult to uniquely differentiate between them. PolV sequences cannot be aligned to PolIV structure in a structural alignment, demonstrating the need to alternative methods.

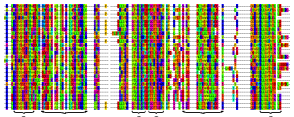


**Figure 6.** Alignment with T-Coffee showing structural conservation in LF. (Subset of alignment shown)

2. Little finger region demonstrates wide range diversity , but mostly conserved secondary structure. The variability in length for each subunit confounds efforts to build a one-size-fits all solution to predicting the LF in all sequences.

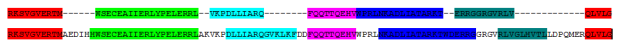

**Figure 7.** Alignment between little finger region previously annotated by hand and little finger annotated automatically. E. coli DNA Pol IV sequence.

3. The Little finger finder currently works well for E. coli, finding the start and end beta helices with 100% accuracy as shown in Fig. 7.
4. Additional bridge found in human Pol η appears to share high similarity with a bacteriophage which infects members of the Saccharomyces genus. This bridge structure appears to exist in most eukaryotic Pol η.

## FUTURE DIRECTIONS

**Improved little finger detection via secondary structure prediction.** The current algorithm relies on single amino acid position weighting. By using secondary structure detection, the algorithm can score regions known to fit the secondary structure requirements, rather than needing to consider all possible regions. This will lead to greater efficiency and higher accuracy.

**Usage of structural information in alignments.** T-Coffee is an alignment suite that can use structural information to align sequences. We have already begun examining its use (**Fig 6**) with promising results. Better alignments yield more reliable phylogenetic data for determining relationships.

**Amino acid markers for PolIV membership.** The amino acids detailed in **Table 1** were selected based on biochemical importance. We will test the pre-LF region, by position, for amino acids which may be markers for either subclasses or PolIV membership. The results may include the positions described above, but may also yield novel positions of interest which can be examined for biochemical significance.

## ACKNOWLEDGEMENTS

We would like to thank Dr. Gary Benson and Dr. Ed Loechler from Boston University, Levy Vargas and Che Martin from the Department of Biology, Hunter College, City University of New York

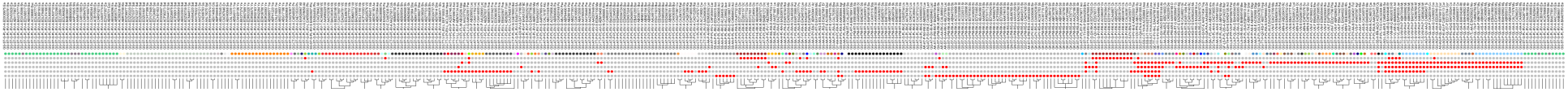

**Figure X.** A fragment of **mrbayes** phylogenetic tree showing all 448 sequences (the labels and branches have been truncated). Colored labels indicate amino acids on several positions of interest according to a template, where each dot corresponds to a bolded component:
G30**G31**-S40T41-**[GP]75**-**P99**X100-**R240**X241-**[^KRQ]250 Genus species**. Color-coding reveals patterns using ● for agreement with template and ● for disagreement. Other colors are used to distinguish genera in the last position.