

# Fingerprint

## *Visual depiction of variation in multiple sequence alignments*

Melanie Lou and G.B. Golding

Department of Biology  
McMaster University

# Outline

- **Web Application: Fingerprint**
- Types of Fingerprints
- Application of the Fingerprint: Lepidoptera
- Why you want to use the Fingerprint!

# Motivation

- Barcode of Life Database (BOLD) tools required for analyzing and displaying COI data effectively
- Various graphical multiple alignment *editors*: ClustalX, Seaview, etc.
- Variety of multiple alignment *shading* programs: BoxShade, T<sub>E</sub>XShade, etc.
- Drawbacks
  - Download and installation, complicated documentation, fees, limits on the number of sequences
  - Focus on sequence-by-sequence representations and not alignment overviews

# Fingerprint

[Home](#)[Documentation](#)[Updates](#)[Contact](#)

## Table of Contents

- [Step 1: Input](#)
- [Step 2: Types of Fingerprints](#)
- [Step 3: Fingerprint Display Options](#)
- [Step 4: Submit/Reset](#)

## Input

Please enter the path to a **fasta** file:

## Types of Fingerprints

### Nucleotides

- Composition
- Heterogeneity
- Identity
- Variability

Scale: Lowest Variability  to Maximum Variability

Max variable sites:  Black  White

- Heterozygosity
- $d_N/d_S$  Ratios
- Nucleotide Diversity

### Amino Acids

- Composition
- Heterogeneity
- Identity
- Variability

Scale: Lowest Variability

Max variable sites:  Bl

OR

- Heterozygosity
- Charges
  - Acidic
  - Basic
- Hydrophobicity
- Solvent Accessibility
- Structure

## Fingerprint Display Options

### Residue Positions

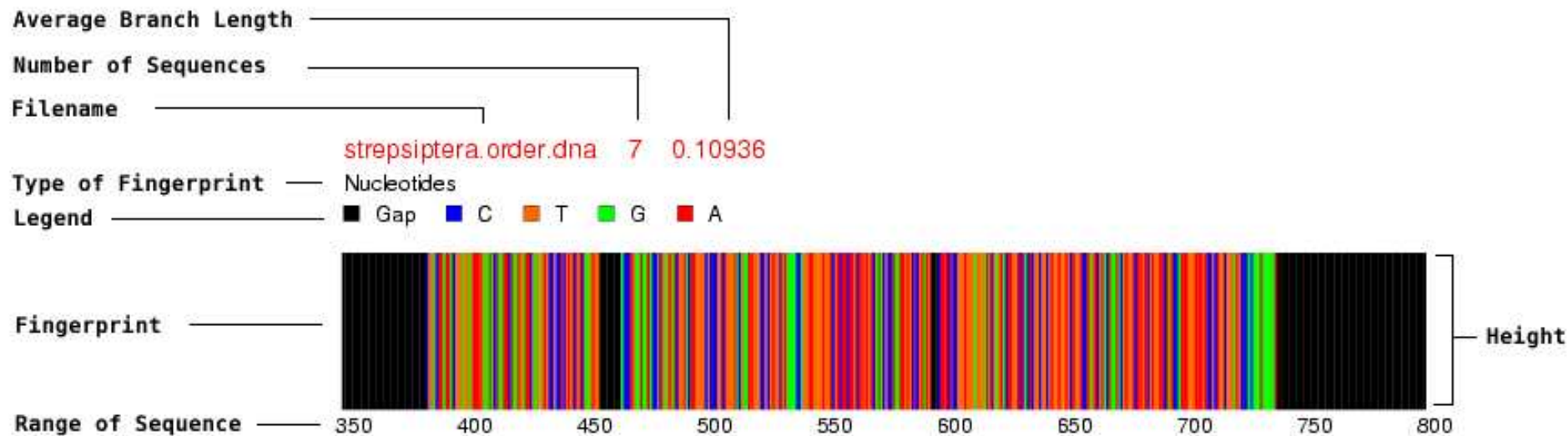
**Default:** first residue set to 1 and last residue set to position of last residue in aligned sequence set

### First and Last Residue Position

FVNQHLCGSHLVEALYLVCGERGFFYTPKAGIVEQCCTGVCSLYQLENYCN  
FVNQHLCGSHLVEALYLVCGERGFFYTPKTGIVEQCCTGVCSLYQLENYCN  
FVNQHLCGSHLVEALYLVCGERGFFYTPKSGIVEQCCTSICSLYQLENYCN  
FVBQHLCGSHLVEALYLVCGERGFFYTPKSGIVDQCCTSICSLYQLENYCN  
FVKQHLCGPHLVEALYLVCGERGFFYTPKSGIVDQCCTSICSLYQLENYCN  
FVNQHLCGSHLVEALYLVCGNDGFFYRPAKAGIVDQCCTGVCSLYQLQNYCN  
APNQRLCGSHLVEALFLICGERGFYYSRSGIVEQCCEINTCSLYQLENYCN  
YVGRLCGSQLVDTLYSVCKHRGF.YRPSEGIVDQCCTNICSRNQLLTYCN

# Fingerprint

A *fingerprint* is a horizontal bar made up of coloured or gray-scale vertical lines representing an **overview** of a desired feature in a sequence or in a set of aligned sequences.



# Outline

- Web Application: Fingerprint
- **Types of Fingerprints**
- Application of the Fingerprint: Lepidoptera
- Why you want to use the Fingerprint!

# Types of *fingerprints*

<b>Nucleotides</b>	<b>Amino Acids</b>
Composition	
Heterogeneity	
Identity	
Variability	
Heterozygosity	
$d_N/d_S$ Ratio	Charges
	Hydrophobicity
Nucleotide Diversity	Solvent Accessibility
	Structure

# Composition

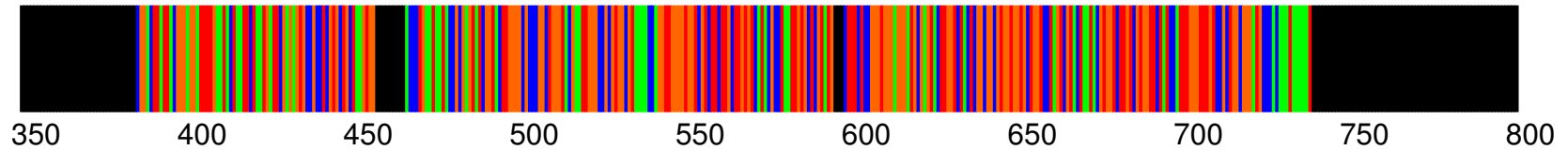
Depicts the residues encoded by the a sequence or an aligned sequence set

## Nucleotides

strepsiptera.order.dna 7 0.10936

Nucleotides

■ Gap ■ C ■ T ■ G ■ A

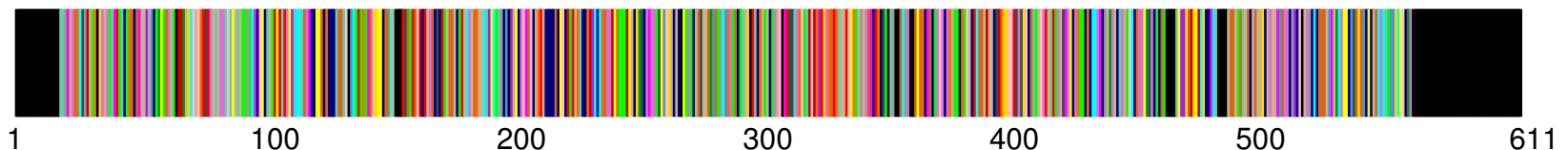


## Amino Acids

hemiptera.order.aa 1121 0.06226

Amino Acids

■ Gap ■ M ■ K ■ W ■ I ■ F ■ S ■ T ■ N ■ H ■ D ■ G ■ L  
■ Y ■ A ■ R ■ E ■ Q ■ P ■ V



# Heterogeneity

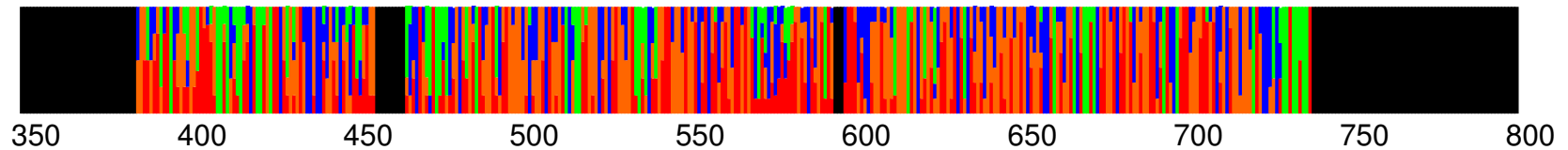
Each residue at a site is represented such that it reflects its frequency at that site

## Nucleotides

strepsiptera.order.dna 7 0.10936

Heterogeneity

■ Gap ■ T ■ C ■ A ■ G

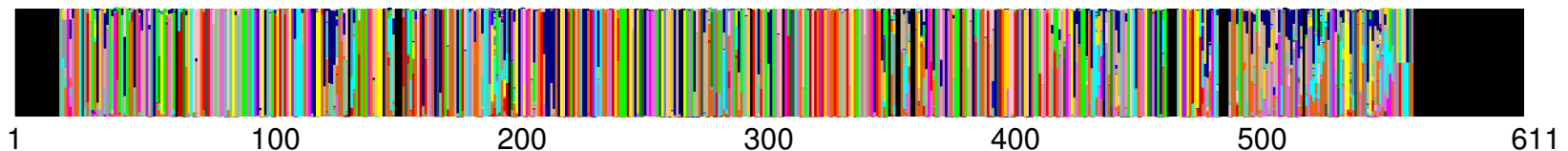


## Amino Acids

hemiptera.order.aa 1121 0.06226

Heterogeneity

■ Gap ■ M ■ N ■ K ■ Y ■ T ■ I ■ V ■ W ■ A ■ F ■ L ■ S  
■ H ■ D ■ C ■ G ■ P ■ R ■ E ■ Q



# Identity

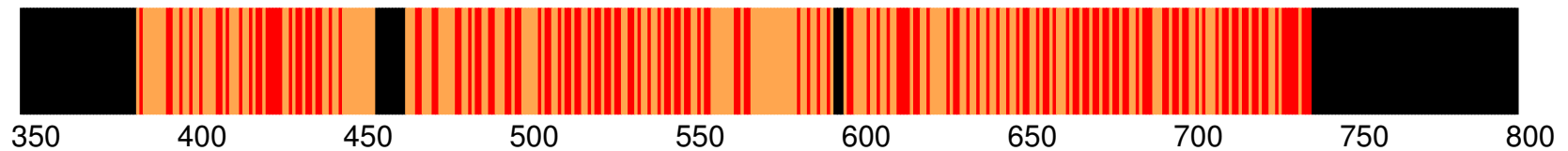
Differentiates between invariant and variant sites

## Nucleotides

strepsiptera.order.dna 7 0.10936

Identity

■ Gap ■ Varied ■ Constant

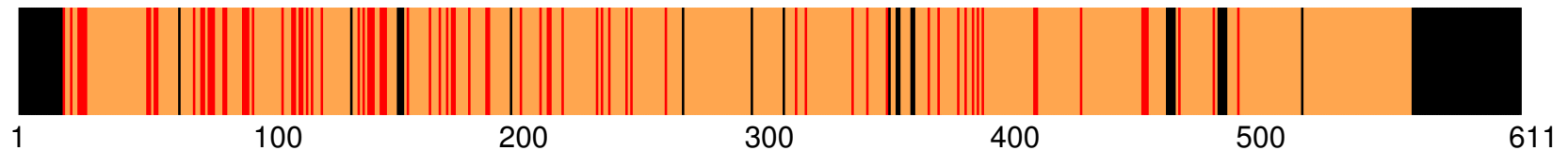


## Amino Acids

hemiptera.order.aa 1121 0.06226

Identity

■ Gap ■ Constant ■ Varied



# Variability

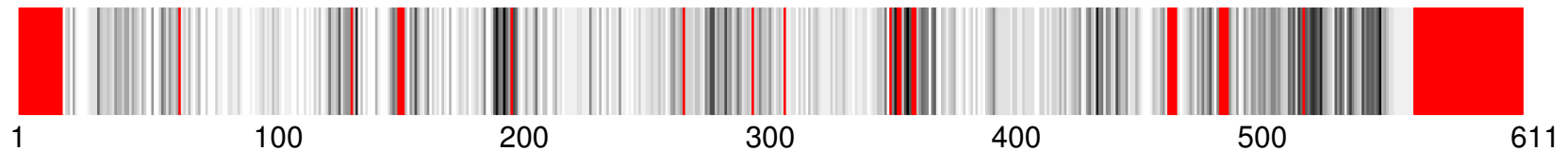
Variability is quantified as the number of possible residues occurring at a site and is coloured accordingly. Sites of maximum variability can be coloured:

## Black

hemiptera.order.aa 1121 0.06226

Variability

■ Gap □ 1 Variant(s) ■ 18 Variants

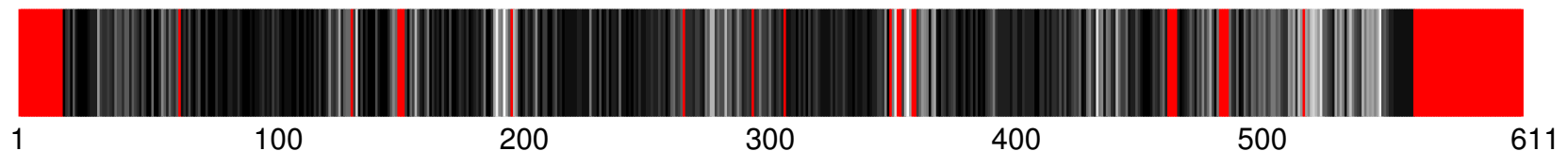


## White

hemiptera.order.aa 1121 0.06226

Variability

■ Gap ■ 1 Variant(s) □ 18 Variants



# Heterozygosity

$$1 - \sum_{i=1}^m x_i^2$$

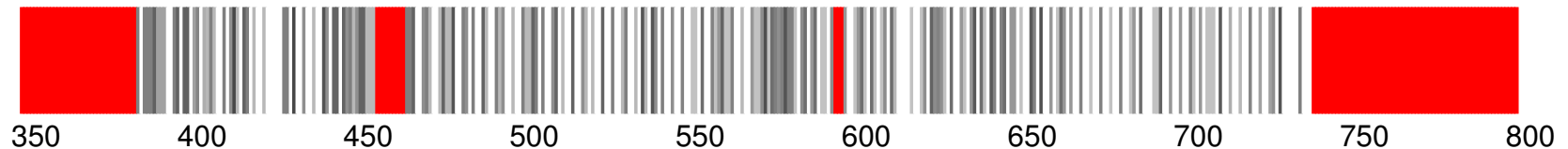
Highly variant sites possess high heterozygosity values

## Nucleotides

strepsiptera.order.dna 7 0.10936

Heterozygosity

■ Gap □ 1 (Zero Heterozygosity) ■ 0 (High Heterozygosity)

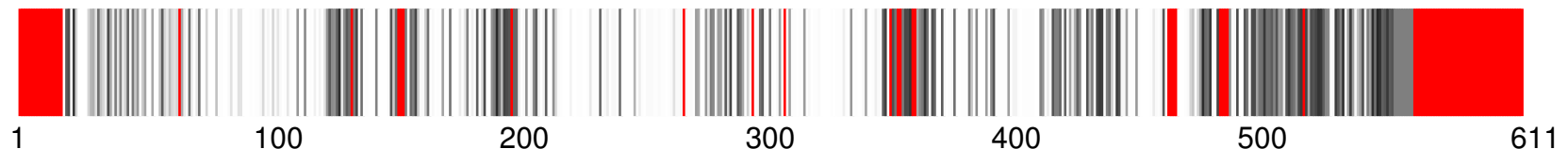


## Amino Acids

hemiptera.order.aa 1121 0.06226

Heterozygosity

■ Gap □ 1 (Zero Heterozygosity) ■ 0 (High Heterozygosity)



# Nucleotide diversity

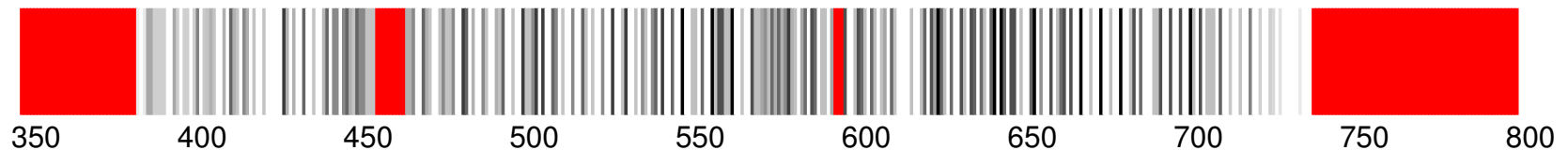
$$\sum_{ij} x_i x_j \pi_{ij}$$

- More suitable for nucleotide data
- Highly variant sites possess high nucleotide diversity values

strepsiptera.order.dna 7 0.10936

Nucleotide Diversity

■ Gap □ Zero diversity ■ Highest diversity



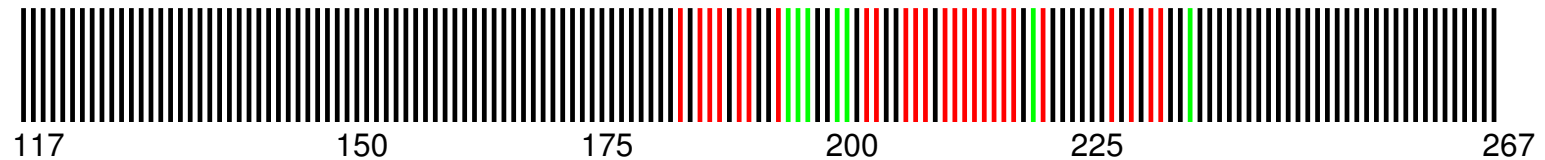
# $d_N/d_S$ Ratio

- Using PAML, the  $d_N/d_S$  ratio is calculated for each codon
- Insight into the type of selective forces that might be in operation

strepsiptera.order.dna 7 0.10936

dN/dS Ratios

■ Undetermined ■ Less than 1 ■ Greater than 1



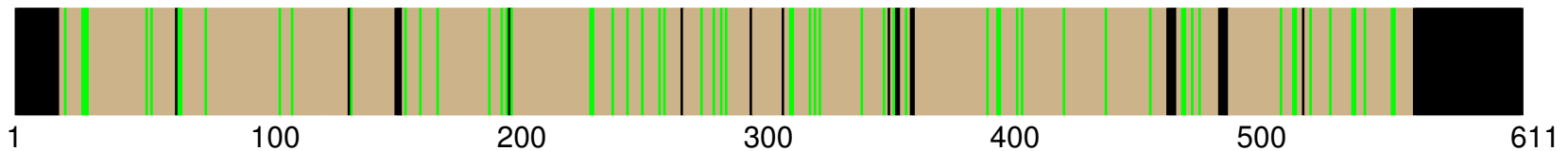
# Charges

## Default

hemiptera.order.aa 1121 0.06226

Charges

■ Gap ■ Uncharged ■ Charged

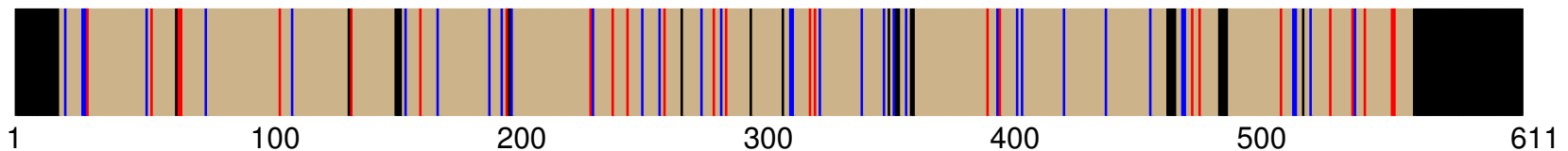


## Acidic or Basic or both

hemiptera.order.aa 1121 0.06226

Charges

■ Gap ■ Uncharged ■ Basic ■ Acidic



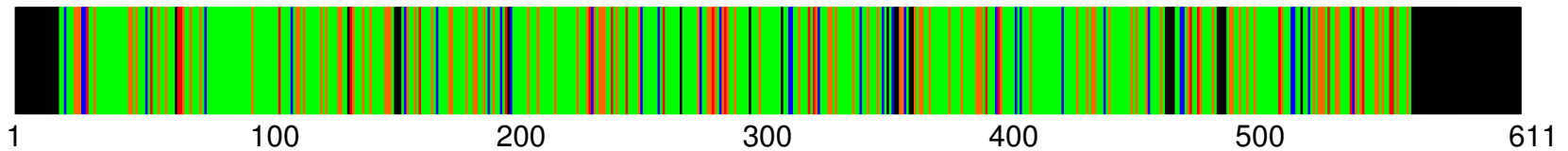
- Identifies charged and uncharged sites
- Charged sites can either be acidic or basic

# Hydrophobicity

hemiptera.order.aa 1121 0.06226

Hydrophobicity

■ Gap ■ Hydrophobic ■ Basic ■ Hydrophilic ■ Acidic



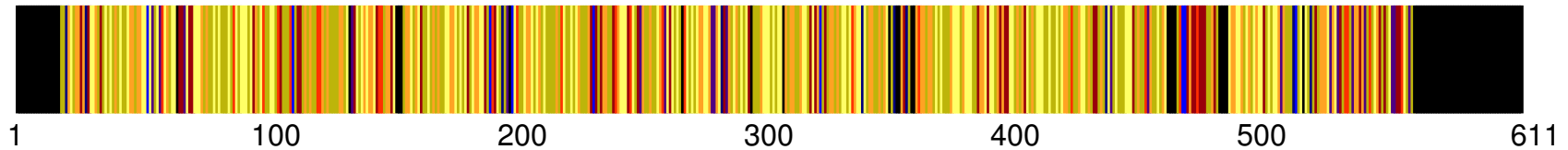
Categorizes sites as acidic, basic, hydrophobic or hydrophilic

# Solvent accessibility

hemiptera.order.aa 1121 0.06226

Solvent Accessibility

■ Gap   □ < 18 Angstrom Sq   ■ 18-27 Angstrom Sq   ■ 28-37 Angstrom Sq  
■ 38-47 Angstrom Sq   ■ 48-57 Angstrom Sq   ■ 58-67 Angstrom Sq   ■ 68-77 Angstrom Sq  
■ 78-97 Angstrom Sq   ■ > 97 Angstrom Sq



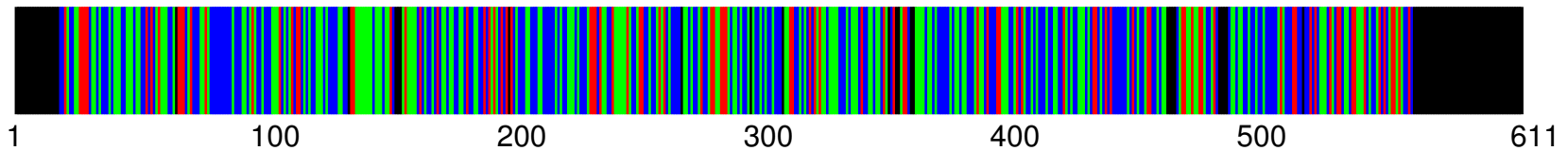
Each amino acid is categorized according to the experimentally determined solvent accessibilities based on the position that such an amino acid is usually found in a folded protein

# Structure

hemiptera.order.aa 1121 0.06226

Structure

■ Gap ■ Internal ■ External ■ Ambivalent



Identifies sites that are localized in the core (internal), on the surface (external) or neither of a globular protein

# Outline

- Web Application: Fingerprint
- Types of Fingerprints
- Application of the Fingerprint: Lepidoptera
- Why you want to use the Fingerprint!

# Lepidoptera: Family

- Applied **Fingerprint** to 9195 lepidoptera sequences
- Initially annotated by genus and species designations
- Subsequently grouped by family
- Generated *composition, variability, heterozygosity* and *nucleotide diversity* fingerprints for each family

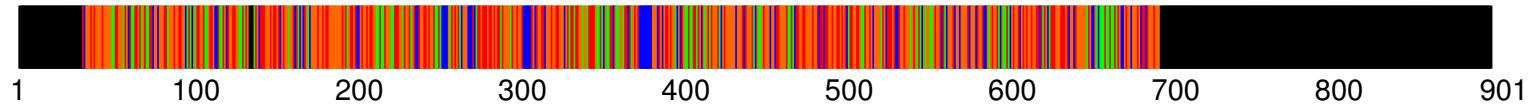
# Lepidoptera: Family

## Nucleotide *Composition* Fingerprints

Crambidae.family 84 0.00325

Nucleotides

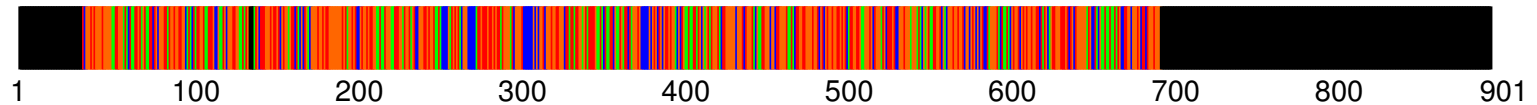
■ Gap ■ A ■ C ■ T ■ G



Gelechiidae.family 122 0.0059

Nucleotides

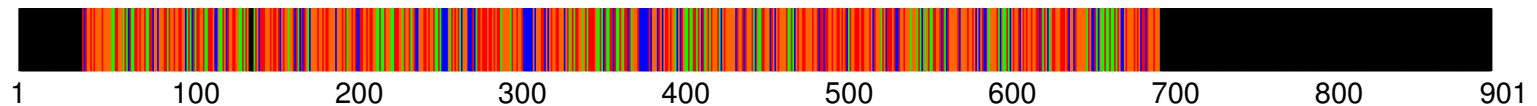
■ Gap ■ A ■ C ■ T ■ G



Nolidae.family 60 0.00363

Nucleotides

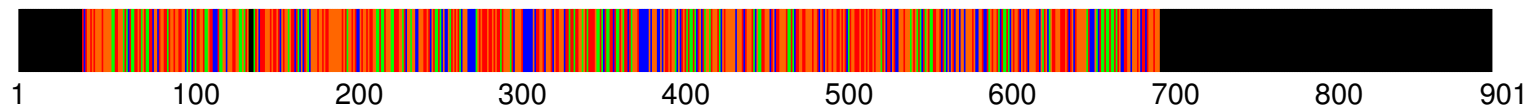
■ Gap ■ A ■ C ■ T ■ G



Pterophoridae.family 64 0.00631

Nucleotides

■ Gap ■ A ■ C ■ T ■ G



# Between families

Despite the composition similarity, the *variability*, *heterozygosity* and *nucleotide diversity* fingerprints revealed distinct patterns of variation between families

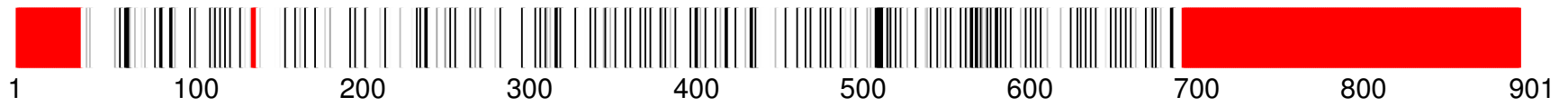
# Between families

## *Nucleotide diversity fingerprints*

**Crambidae.family** 84 0.00325

Nucleotide Diversity

■ Gap □ Zero diversity ■ Highest diversity



**Gelechiidae.family** 122 0.0059

Nucleotide Diversity

■ Gap □ Zero diversity ■ Highest diversity



**Nolidae.family** 60 0.00363

Nucleotide Diversity

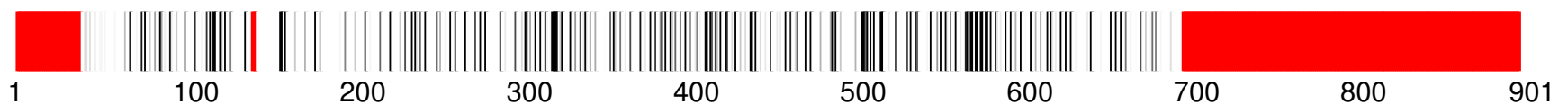
■ Gap □ Zero diversity ■ Highest diversity



**Pterophoridae.family** 64 0.00631

Nucleotide Diversity

■ Gap □ Zero diversity ■ Highest diversity



# Family: Crambidae

Crambidae.family 84 0.00325

Heterozygosity

■ Gap □ 1 (Zero Heterozygosity) ■ 0 (High Heterozygosity)



Variability

■ Gap □ 1 Variant(s) ■ 10 Variants



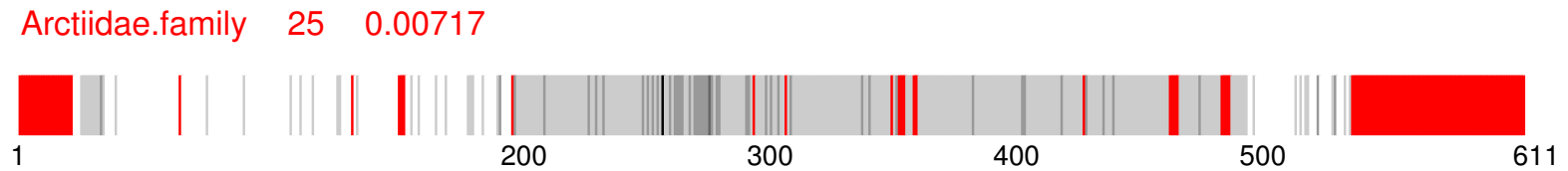
Nucleotide Diversity

■ Gap □ Zero diversity ■ Highest diversity



Positions: 190, 300

# Serendipity!



Ability to identify alignment errors

# Outline

- Web Application: Fingerprint
- Types of Fingerprints
- Application of the Fingerprint: Lepidoptera
- **Why you want to use the Fingerprint!**

# Fingerprint deliverables

- Output is compact, intuitively understandable, and is well suited for providing a quick overview of alignments
- Effectively identify sequence variation
- Applicable to a wide variety of datasets from any sequence data
- Identify mis-alignments
- Provide high quality graphics for presentation
- Globally accessible through major web browsers

Overall, Fingerprint is an effective tool to quickly and intuitively view the similarities, differences, and patterns in multiple sequence alignment.

# Acknowledgements

The Golding lab: G.B. Golding, Ying Fong, Tom Ferguson, Weilong Hao, and other lab members.

This work was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, and other sponsors listed at [www.BOLNET.ca](http://www.BOLNET.ca).

# Thank You

Melanie Lou and G.B. Golding  
McMaster University  
<http://evol.mcmaster.ca/fingerprint>