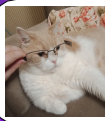


# The mysteries of the Histidine Triad Protein Family in *Streptococcus*

Anna Lybanskaya, Maria Novikova, Yana Savchenko  
 Olga Bochkareva, Natalia Dranenko, Vera Emelianenko, Aygul Nasibullina, Alexander Chistyakov, Ukron TinArden

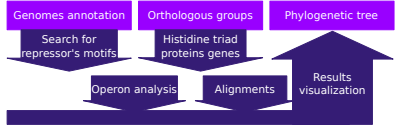


## Introduction

Histidine Triad Proteins (Hit Proteins) is a family of proteins located on the bacterial outer membrane, performing various functions. In some species of *Streptococcus* they are also involved in pathogenicity. Up to 4 different representatives of this protein family can be found in one genome; however, their structure and functions are significantly understudied. During the school, we found Hit proteins genes in *Streptococcus* genomes available in GenBank, aligned their sequences, analyzed the operon structure, and predicted the binding sites of the zinc repressor, associated with those operons.

## Methods

PanAcoT: preliminary analysis; HMMer: search for genes coding Hit-proteins; z-scan: predicting binding sites of zinc repressor; MUSCLE: alignments; iTOL: tree visualization.



## Table of symbols

- Zn + Zn -
- Large length
- Medium length
- Small length

Species with potential phase variations and the number of strains from 4 to 16

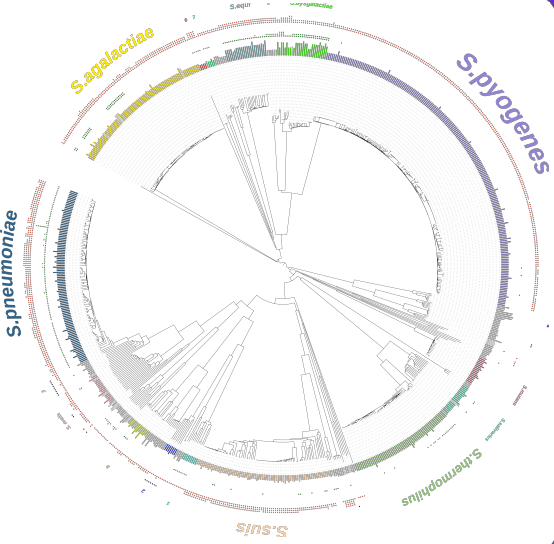
- 1- *S. gordoni*
- 2- *S. sanguinis*
- 3- *S. mitis*
- 4- *S. galloyticus*
- 5- *S. suis*
- 6- *S. iniae*
- 7- *S. uberis*
- 8- *S. anginosus*

Species are labeled with the same color; those species that have > 15 strains also have text labels. Filled empty circled represent Hit proteins encoding genes that have/don't have zinc repressor binding site.

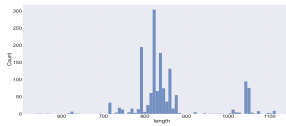
Two clusters are seen on the tree: species that have 0-3 *pht* per strain and species that have 2-7 *pht* per strain.

The first cluster (0-1 *pht*) is mostly represented by non-pathogens (some are even used as probiotics) and opportunists (some can cause caries), such as *S. thermophilus*, *S. salivarius*, *S. mutans*, etc. These species either don't have *pht* or have only one short copy without a zinc repressor binding site.

The second cluster is mostly represented by pathogens such as *S. pyogenes*, *S. pneumoniae*, *S. agalactiae*, etc. They often have *pht* genes of a medium length (1512 findings, 41 species), seldom have long *pht* genes (23 findings, 8 species), and also have short *pht* genes (324 findings, 27 species).

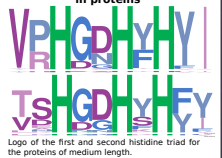


## Length of Hit proteins



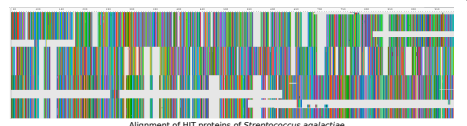
Hit proteins that we found using HMMer have very different lengths. The histogram of the lengths distribution shows that most of the proteins are 500-1150 aminoacids long. The rest can be divided into short (<500 aa long) and long proteins (>1150 aa long).

## Histidine triad's structures in proteins

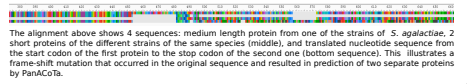


Logo of the first and second histidine triad for the proteins of medium length.

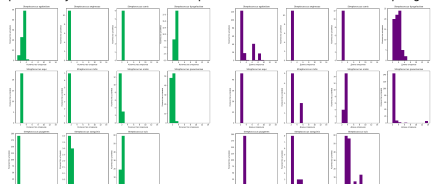
The first two histidine triads in the proteins of medium length are the most conservative. We noticed that the amount of histidine triads is correlated with the length of the proteins: long Hit proteins can have up to 6 and more histidine triads, while the proteins of medium lengths usually have 3-4 triads. In some cases, short Hit proteins found by HMMer didn't have histidine triads at all. For example, in *S. thermophilus* all proteins found by HMMer are short, don't have histidine triads and zinc repressor binding site. Given that *S. thermophilus* is not pathogenic, we suppose that the presence of Hit proteins in this species is an artifact.



We aligned HIT proteins of (short, long and medium altogether) for every *Streptococcus* species in our dataset. Inspecting the alignments manually, we noticed that some short proteins are in fact fragments of longer proteins, for example in *S. agalactiae*. This is consistent with the HIT proteins distribution across the phylogenetic tree. ORFs of these short proteins are located one after another.



From the literature we know that some genes encoding for Hit proteins are regulated by zinc repressors. We searched for potential binding sites of zinc repressors in the upstream of *pht* genes from our dataset. Most of the *pht* genes indeed have zinc repressors binding sites and are presumably under control of zinc repressors. We have also noticed that *pht* genes group in operons.



Inside one species the number of operons with *pht* genes is more or less stable - about 2 operons per species. In contrast, the length of the operons is more variable. In a lot of species, these operons have roughly the same length and include only *pht* genes and genes of zinc import (*znuA*), but exceptions also occur. In *S. mitis*, operons with *pht* genes can have either 2 or 5 genes, but they always consist only of *pht* and *znuA* genes. In *S. agalactiae*, *S. disgalactiae*, and *S. suis*, operon lengths are more variable.

Some of the operons from *S. pneumoniae* include up to 13 genes, but we think that this high amount is the result of pseudogenes remnants.

