

# Connecting Sequence Data to Virulence Factors in Streptococcus Genomes

---

Mathematics and Computer Science Division

**About Argonne National Laboratory**

Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, see [www.anl.gov](http://www.anl.gov).

**DOCUMENT AVAILABILITY**

**Online Access:** U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via DOE's SciTech Connect (<http://www.osti.gov/scitech/>)

**Reports not in digital format may be purchased by the public from the National Technical Information Service (NTIS):**

U.S. Department of Commerce  
National Technical Information Service  
5301 Shawnee Rd  
Alexandria, VA 22312  
**[www.ntis.gov](http://www.ntis.gov)**  
Phone: (800) 553-NTIS (6847) or (703) 605-6000  
Fax: (703) 605-6900  
Email: **[orders@ntis.gov](mailto:orders@ntis.gov)**

**Reports not in digital format are available to DOE and DOE contractors from the Office of Scientific and Technical Information (OSTI):**

U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831-0062  
**[www.osti.gov](http://www.osti.gov)**  
Phone: (865) 576-8401  
Fax: (865) 576-5728  
Email: **[reports@osti.gov](mailto:reports@osti.gov)**

**Disclaimer**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor UChicago Argonne, LLC, nor any of their employees or officers, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, Argonne National Laboratory, or UChicago Argonne, LLC.

## **Connecting Sequence Data to Virulence Factors in Streptococcus Genomes**

---

prepared by

Jeyashree Alagarsamy  
Madurai Kamaraj University, Madurai, Tamil Nadu, India

Ramya Seetharaman  
Chandramouli Sivanandham  
Anna University, Chennai, Tamil Nadu, India

Karthickeyan Chella Krishnan  
Department of Molecular Genetics, University of Cincinnati

Ross Overbeek  
Mathematics and Computer Science Division, Argonne National Laboratory

March 31, 2014

## Connecting Sequence Data to Virulence Factors in *Streptococcus* Genomes

The SEED Project ([http://www.theseed.org/index.php/Main\\_Page](http://www.theseed.org/index.php/Main_Page)) was started over a decade ago to focus on creating more accurate annotations of prokaryotic genomes, along with the tools needed to support such an effort. A number of research teams have participated, and numerous projects were based on the evolving technology that was cooperatively built. Some of these teams sought funding to support comparative analysis of pathogens, and a number did successfully acquire funding from NSF, NIH, and DOE.

While it is often asserted that we cannot afford to support manual annotation efforts, we counter with the following simple argument:

1. Accurate automated annotations will directly impact the value of the hundreds of thousands of genomes that will be sequenced during the next few years; and
2. The quality of automated annotations is directly related to the availability of accurately annotated reference genomes.

While it is true that most manual annotation efforts could have achieved far higher efficiency, it is also true that extraction of value from newly sequenced genomes will depend heavily on a carefully curated set of *subsystems* built upon the annotations of a set of reference genomes.

Ultimately, accurate annotations need to be connected to the supporting literature. Since a number of groups utilize the SEED technology for research projects and since we were initiating a pilot project relating to *Streptococcus pyogenes*, we decided to gather literature relating to virulence factors in *Streptococcus* species, to relate papers to specific genes in the pubSEED database, and to make the connections we derived freely available. We have not connected all the articles to specific genes because, in a few cases, the actual strains that were experimentally characterized have not even been sequenced (in other cases, they probably have been sequenced, but we could not reliably make the needed connection).

The largest impact of this effort will probably be via the connections from genes to literature as supported in pubSEED. Our intent, however, is to provide the data in a form that makes it conveniently accessible to any annotation effort.

We include here a summary of the *Streptococcus* species for which data was gathered.

## **Summary of the *Streptococcus* Species Analyzed**

Bacteria belonging to the genus *Streptococcus* are all Gram-positive and spherical.

***Streptococcus pneumoniae*** is one of the major human pathogenic bacteria from the Genus *Streptococcus*. *S. pneumoniae* is a major cause of bacterial Meningitis. It also causes various pneumococcal diseases such as acute sinusitis, otitis media, conjunctivitis, bacteremia, sepsis, osteomyelitis, septic arthritis, endocarditis, peritonitis, pericarditis, cellulitis, and brain abscess. *S. pneumoniae* resides on mucosal surfaces in nasopharynx but can spread to other location causing infections in immune-compromised people and children.

***Streptococcus iniae*** is a major fish pathogen and occasionally infects humans. *S. iniae* causes meningoencephalitis, septicemia, and central nervous system damage in fishes, and causes meningitis, endocarditis, osteomyelitis, and septic arthritis in immune-compromised people. The site of *S. iniae* infection varies from species to species in case of fishes, while source of infection is not determined in humans.

***Streptococcus mitis*** is closely related to *S. pneumoniae*. *S. mitis* are not usually pathogenic, but commonly cause bacterial endocarditis. It resides in hard surfaces in the oral cavity such as dental hard tissues as well as mucous membranes and is part of the oral flora.

***Streptococcus pyogenes* a.k.a Group A Streptococcus (GAS)** is associated with wide variety of disease conditions in humans. The majority of GAS diseases are associated with skin and pharyngeal mucosa; however, occasionally GAS also causes life-threatening conditions such as necrotizing fasciitis and toxic shock syndrome. The annual death rate due to GAS infection is estimated to about 150,000 worldwide.

***Streptococcus parauberis*** is a known pathogenic bacterium to cause two economically important diseases: mastitis in dairy cattle and streptococcosis in fish. Only limited research has been done on *S. parauberis*.

***Streptococcus uberis*** is responsible for a significant proportion of bovine mastitis in commercial dairy herds. It colonizes body sites of the cow including the gut, genital tract and mammary gland.

***Streptococcus agalactiae*** a.k.a. **Group B Streptococcus (GBS)** is a well-known causative agent for neonatal sepsis. From a veterinary point of view, this bacterium is considered one of the major causes of bovine intramammary infections, particularly in North America, and is a source of economic loss for the industry.

***Streptococcus suis*** is an important pathogen in pigs. *S. suis* is the name assigned to streptococci that were formally called Lancefield groups R, S, and T. *S. suis* can also be communicated from pigs to humans. Human infections include meningitis, septicemia, endocarditis, and deafness.

***Streptococcus mutans*** is found in the human oral cavity. *S. mutans* is considered to be the most cariogenic of all the streptococci. It can thrive in high temperature such as 18-40°C, which helps it to metabolize different kinds of carbohydrates, creating acidic environment in the mouth causing tooth decay. *S. mutans* is implicated in the pathogenesis of certain cardiovascular diseases and is the most prevalent bacterial species detected in extirpated heart valve tissues, as well as in atheromatous plaques.

***Streptococcus dysgalactiae* subsp. *equisimilis*** belongs to Lancefield groups C and G. It causes infections in cows, and humans. It is found in the oral cavity of humans. It causes invasive streptococcal infections such as streptococcal toxic shock syndrome and mastitis.

***Streptococcus gallolyticus* subsp. *pasteurianus*** is a Lancefield group D streptococcus, formerly known as *Streptococcus bovis* biotype II. *S. pasteurianus* are isolated from blood cultures of patients with colonic cancer. It is also reported to cause adult meningitis.

***Streptococcus gallolyticus* subsp. *Gallolyticus***, formerly referred to as *Streptococcus bovis* biotype I, is a member of group D streptococci. It is found in the alimentary tract of cows, sheep, and humans. It is responsible for endocarditis in a patient suffering from colon

cancer, but knowledge about the virulence factors is limited, and its pathogenesis is not fully understood.

## **Results**

The virulence factors for all the species were identified from the experimental proofs available in the research papers for different strains used in this analysis. We chose the virulence factors based on the experimental data where wild type strain and mutant strain (virulence factor deleted from the strain) is compared and data analyzed to identify the function of the virulence factor.

However, we encountered few issues while collecting the data.

1. We were unable to find the Gene ID/Genome ID/PubSEED ID/NCBI ID (their roles given in PubSEED/NCBI), even though we were able to find the experimental proof. The reason may be that the strain has not yet been deposited into NCBI/PubSEED or it has not yet been sequenced.
2. In some cases, virulence factors for the given strains were not experimentally tested but, instead, were characterized based on sequence comparison and alignment. In brief, the query strain is compared with a different strain or a species harboring well-known virulence factors and is in some way related to the query strain; based on this alignment and sequence comparison, virulence factors are proposed. These are theoretical data, but we cannot ignore them because these might be virulence factors. Further experiments are needed in order to determine whether these are actually virulent or not.
3. The last issue was a slight variant of the second. Certain strains are proposed to harbor putative virulence factors but do not have any experimental verification associated with them. Instead they were compared with other well-known virulence factors based on sequence alignments. These virulence factors are mentioned as hypothetical genes in both NCBI and PubSEED and therefore warrant further experimentation in order to identify the function of these putative virulence factors.









**Mathematics and Computer Science Division**

Argonne National Laboratory  
9700 South Cass Avenue, Bldg. 240  
Argonne, IL 60439-4847

[www.anl.gov](http://www.anl.gov)



Argonne National Laboratory is a U.S. Department of Energy  
laboratory managed by UChicago Argonne, LLC