FluHunter Database

A close up of a document

Description automatically generated

A close up of a text

Description automatically generated

A screenshot of a computer

Description automatically generatedA close-up of a document

Description automatically generated4 gruppi:

* Marker
* Literature
* Mutation
* Segment Data

In FluHunter’s database it is possible to distinguish four main area of interest in the domain of the Influenza virus, plus the table Subtype, common to the entire domain.

The table Subtype it is used to identify a clade of Influenza virus, and is a specification to the Influenza type. Subtypes of Influenza A, considered in this document, are defined by the combination of the antigenic viral proteins haemagglutinin (H) and neuraminidase (N) in the viral envelope.   
The table itself is composed by an id, used to identify each subtype univocally, and by the name of the subtype itself (ie, “H5N1”, “H3N2”, etc).

The first area of interest has to do with the collection, storage and structuring of the available information on actual Influenza samples from online repositories such as GISAID. This area could be generally referred to as Samples Data and it contains two main tables, the Isolate and Segments tables, and a secondary table, Location, to store more conviniently location data, reducing data redundancy in the database and helping maintaing database normality.

Going into more detail, table Isolate is used to encompass a single sample of Flu virus, as describing the entire genetic material taken from a single host on a specific date, and it is composed of the following fields:

* isolate\_id (INTEGER, PRIMARY KEY, AUTO\_INCREMENT): A unique identifier for each collected sample, it is present as a way to uniform the tables in the schema, as isolates in GISAID are identified by their “EPI” value, thus making an id redundant in practice;
* isolate\_epi (TEXT, UNIQUE): The “EPI” value is used by GISAID as a way to uniquely identify each sample, it comes in the format “EPI\_???” ;
* subtype\_id (INTEGER): An id referencing the Subtype table, it is used to identify which subtype the isolate is from, to correctly indentify information about the isolate itself, for example and isolate of H5N1 subtype will have a subtype\_id of 22.
* host (TEXT): This field describes the host the genetic material of the isolate was taken from. It is saved as a text field instead of an id field referencing a separate table because the Isolate table is the only one making use of this field and further information on hosts was not available nor interesting for this work, thus making a separate table not optimal.
* collection\_date (TEXT): The date on which the Isolate was collected from the host. They come in various formats since date collection is not regulated on GISAID, examples range from the sole indication of the extended year, ie “1959”, to dates formatted as “YYYY-MM-DD”.
* location\_id (INTEGER): Ad id referencing the Location table, used to identify the location in which the sample was collected. Being locations composite fields it was decided to create a table to better represent the information, thus making an id referencing necessary.

Table Location, as previously said is used to identify a particular location in the world. The table itself is composed of four fields: an id to uniquely identify each location; the region which broadly indicates the continent (for example Europe or Africa), used to distinguish between locations with the same name in different places; the state field which can be equipared to the name of a country (e.g: Italy, United Kingdom); and finally city, for example Milan. It is to be noted that data was garanteed for regions only, thus some Locations are missing of the state/city fields, which are initialized to Null values. For our analysis locations with missing fields are only considered in queries that use the non-null fields.

Finally table Segment stores the metadata associated with a particular segment type or “protein” (e.g. “HA”, “NA”, etc) whose genetic material was extracted from a single Isolate. In this sense, for each Isolate, the distinct proteins are treated as separate entities with different metadata. It is to be noted that, although up to all eight segments of an isolate can be present in the database, this is far from the norm, with a average of five segment for isolate across the current data.

Going into the specifics, table Segment encompasses the following fields:

* segment\_id (INTEGER, PRIMARY KEY, AUTO\_INCREMENT): A unique identifier for each collected sample, for each isolate.
* isolate\_epi (TEXT): This field is used to reference the Isolate table as a way to indentify each Segment to the Isolate it comes from. It has thus the same format as the isolate\_epi field in the Isolate table and the connection between the two is automatically made.
* segment\_type (TEXT): The name of the protein the Segment represents, e.g. “HA” in the case of an Hemagglutinin protein Segment.
* segment\_epi string
* virus\_name string
* epi\_virus\_name (TEXT): a combination of of segment\_epi and virus\_name, it is used to uniquely identify a Segment in the GISAID dataset.

The second area of interest is that of Mutations Data,