**ER-Model**

**A close up of a map

Description automatically generated**

|  |  |
| --- | --- |
| **Legend:** | |
| pk | Primary key |
| fk | Foreign key |
| cpk | Composite primary key |
| A close up of a logo  Description automatically generated | Entity.  Attributes are noted inside the entity. Example: **seqid** is a primary key attribute here of entity **sequence** having datatype **int**. |
| A picture containing sitting, table  Description automatically generated | Relationship between entities |
|  | One to one association |
| A close up of a device  Description automatically generated | One to many association |
| A close up of a device  Description automatically generated | Many to many association |
|  | Partly optional association |
| Items in bold | Primary/composite keys. |
| Italicized items | Foreign keys |

Description of ERD components:

|  |  |  |  |
| --- | --- | --- | --- |
| Entity | Key type | Field | Notes |
| Blastn\_output | Candidate keys/composite primary keys | seqNo and configNo | - |
| Foreign keys | sseqid | Corresponds to primary key of subject\_sequence entity |
| configNo | Corresponds to primary key of configuration entity |
| seqNo | Corresponds to primary key of input\_sequence entity |
| Calc\_fitness | Candidate keys/composite primary keys | seqNo and configNo | - |
| Foreign keys | configNo | Corresponds to primary key of configuration entity |
| seqNo | Corresponds to primary key of input\_sequence entity |
| Configuration | Primary key | configNo | - |
| Other\_calculated\_values | Primary key | recordNo | This is a temporary table |

Notes:

* All the entities above can be represented as individual tables.
* I**nput\_sequence** and **subject\_sequence** can be read from flat files at the moment while performing blastn queries, thus there is no need to load them as tables into the database.
* We implemented an RDBMS for our project. Our data is structured and there is a need for integrity across data.
* The tables/entities above have been normalized to 1-normal form. Inducing the composite key in two of the tables ensures data integrity.
* The ERD above can be deployed as *logical schema* for further analysis of blastn output data.

**Table 1:**

Table name: **configuration**

This table bears the varying configurations that are being used for the experiment.

Total no of configuration sets: 3000

Here, a single configuration set is a combination of 7 different independent configurations.

Table header/field names (from left to right):

1. **configNo**
2. **dust**
3. **lcase\_masking**
4. **soft\_masking**
5. **ungapped**
6. **xdrop\_gap**
7. **xdrop\_gap\_final**
8. **xdrop\_ungap**

**Table structure: configuration**

**A screenshot of a cell phone

Description automatically generated**

|  |  |
| --- | --- |
| **Configuration name** | **Possible values** |
| dust | yes, “20 64 1”, no |
| lcase\_masking | true, false |
| soft\_masking | true, false |
| ungapped | true, false |
| xdrop\_gap | 0, 0.1, 0.5, 30, 100 |
| xdrop\_gap\_final | 0, 0.1, 0.5, 10, 100 |
| xdrop\_ungap | 0, 0.1, 0.5, 20, 100 |

Field 1 is the configuration no/ the identifier for a single configuration set.

Field 2-8 are the values of the configurations for any configuration set in blastn, that we are using for the experiment.

The image below depicts a snapshot of the table **configurations**:

A screenshot of a cell phone

Description automatically generated

**Figure 1: View schema for the table ‘configurations’**

* In the image, we find the exact values pertaining to each configuration number.
* For example: configuration no 1 above has dust= “yes”, lcase\_masking=”true”, soft\_masking=”true”, ungapped=”true”, xdrop\_gap=0.0, xdrop\_gap\_final=0.0, xdrop\_ungap=0.0

**Table 2:**

Table name: **blastn\_output**

This table bears the blastn output data, generated after executing each blastn query.

Total no of entries in this table = 96875728 rows.

The table does not contain entries for sequences which returned no hits against a configuration.

Table header/field names (from left to right):

1. **SeqNo**
2. **ConfigNo**
3. **hitno**
4. **qseqid**
5. **sseqid**
6. **pident**
7. **length**
8. **mismatch**
9. **gapopen**
10. **qstart**
11. **qend**
12. **sstart**
13. **send**
14. **evalue**
15. **bitscore**
16. **qhits**

**Table structure: blastn\_output**

A close up of text on a white background

Description automatically generated

Field 1 is the sequence no. Total no of possible sequences= 10000

Field 2 is the configuration no. Total no of possible configurations= 3000

Field 3 is the hit no pertaining to a particular hit for a given sequence under a given configuration.

Field 4 is the query sequence for which we are trying to find a match.

Field 5 is our reference/subject genome sequence id.

Field 6 is the percentage of identical matches (percentage identity).

Field 7 is the alignment length.

Field 8 is the number of mismatches.

Field 9 is the number of gap openings.

Field 10 is the start of alignment in query.

Field 11 is the end of alignment in query.

Field 12 is the start of alignment in subject.

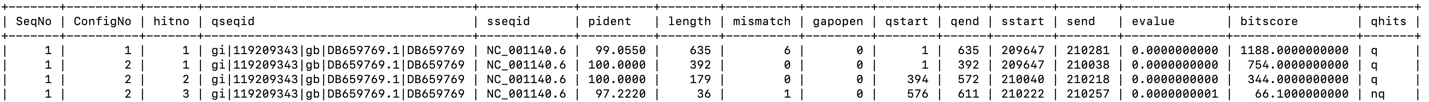
Field 13 is the end of alignment in subject.

Field 14 is the expect value (e-value).

Field 15 is the bitscore.

Field 16 describes if a hit is a quality hit or not. If a hit had pident>=99 and evalue<=0.01 then qhit=’y’ else qhit= ‘n’.

Figure 1 depicts a snapshot of the table **blastn\_output table**



**Figure 2: View schema for the table ‘blastn\_output’**

* The first column depicts the sequence no.
* The second column shows the configuration no.
* The data presented above represents 1 hit for sequence no 1, configuration no 1.
* It also shows 3 rows of hits for sequence no 1, configuration no 2. The no of times a particular configuration no repeats itself for a particular sequence no, denotes the total no of hits the individual sequence has encountered under that given configuration. Column 16 denotes whether the hit found is a quality hit or not.

**Table 3:**

Table name: **fitness**

This table bears the **fitness data**, calculated from the output files generated during after executing each blastn query.

Total rows of data in table = 29.898,740

Data contains values for 10,000 sequences over 3,000 configurations each.

The table does not contain an entry for a sequence which return zero hits under a given configuration.

Sequences for which a particular configuration returned no hit have been filled with value of hits=0 and arbitrary values for all other columns.

Table header/field names (from left to right):

1. **SeqNo**
2. **ConfigNo**
3. **Hits**
4. **Distance**
5. **AvgPercId**
6. **AverageEval**
7. **MedianPercId**
8. **MedianEval**
9. **W\_dist**
10. **W\_eval**
11. **FitnessScore**

**Table structure: fitness**

**A screenshot of a cell phone

Description automatically generated**

Field 1 depicts the sequence no.Total no of possible sequences= 10,000.

Field 2 depicts the configuration no. Total no of possible configurations= 3,000.

Field 3 is the exact number of hits obtained per sequence, per configuration.

Field 4 is the distance of the configuration from the default configuration in blastn (i.e, the no of individual parts in the configuration set that differ from the default configuration set. Here, the configuration set is a combination of 7 different independent configurations).

Field 5 is the average percentage identity across all hits observed per configuration, for a particular sequence.

Field 6 is the average evalue across all hits observed per configuration, for a particular sequence.

Field 7 is the median percentage identity calculated across all hits observed per configuration, for a particular sequence.

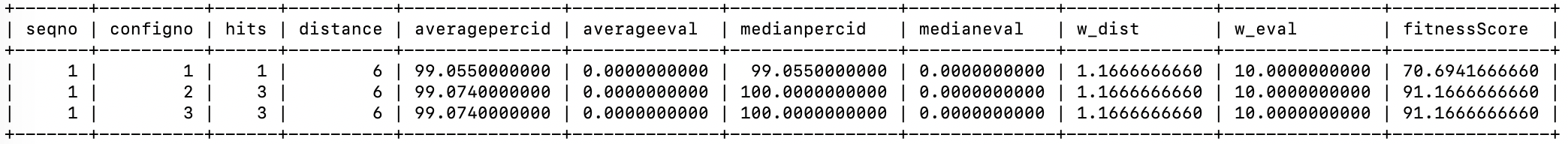
Field 8 is the median eval calculated across all hits observed per configuration, for a particular sequence.

Field 9 is the weighted distance w\_dist, where w\_dist= 10 if distance =0, and w\_dist= 7/distance, if distance is not =0.

Field 10 is the weighted evalue w\_eval, where w\_eval= 10 if evalue=0, and w\_eval =1/median evalue if w\_eval is not=0.

Field 11 is the calculated fitnessScore using our defined formula.

The below image depicts a snapshot of the table **fitness**



**Figure 3: View schema for the table ‘fitness’**

* The data represents results for sequence no 1 and configuration no ranging from 1 to 3. Each row of data depicts the sequence number, configuration number, the calculated fitness scores, the average and median percentage identities, the average and median e-values, hits, the weighted evalue and weighted distance.