File information

Overview

- `*_orig.ct` is the original ct file
- `*_nop.ct` contains the result of removing pseudoknot base pairs from the original ct file. `*_knots.txt` contains a list of the removed base pairs.
- `*_canon.ct` contains the result of removing non-canonical base pairs (pairs that are not Watson-Crick or GU/UG wobble pairs) from `*_nop.ct`.
  - `*_noncanonical.txt` contains a list of the removed base pairs.
- `*_clean.ct` contains the result of removing isolated base pairs (pairs `(i,j)` for which neither pair `(i-1,j+1)` nor `(i+1,j-1)` exists) from `*_canon.ct`.
  - `*_isolated.txt` contains a list of the removed base pairs.

- gtmfe was used to calculate the mfe of the structure containing the base pairs in `*_clean.ct`. Constraints forcing the formation of base pairs were generated based on `*_clean.ct`, but the unpaired bases in `*_clean.ct` were not constrained in any way.
  - The resulting structure is saved as `*_forced.ct`
  - The energy of `*_forced.ct` is saved in the `forced_energy` column.
- gtmfe was used to calculate the mfe of the unconstrained structure.
  - The resulting structure is saved as `*_mfe.ct`.
  - The energy of `*_mfe.ct` is saved in the `mfe_energy` column.

About accession numbers

- Although each row in the database has an `accession` field giving the sequence's accession number, these numbers are not necessarily unique across rows. Some of our sequences are only fragments of the "official" sequence that the NCBI Nucleotide database associates with the given accession number. So we may have multiple sequences in the database that are different portions of this "official" sequence.
- We only use information about the "official" sequence for three fields:
  - `acc_length` gives the length of the official sequence.
  - `seq_start` gives the starting index of our sequence within the official sequence (0-indexed). Its value is NULL when neither our sequence nor the reverse complement of our sequence appears within the official sequence.
  - `seq_stop` gives the ending index + 1 of our sequence within the official sequence (0-indexed). Its value is NULL exactly when `seq_start` is NULL.
- When the header for one of our sequences contains multiple accession numbers, we examine each one, giving priority to numbers that come earlier.
- When our sequence does not appear within the official sequence, but the reverse complement of our sequence does appear within the official sequence... **TODO**

Detailed table information

- **Note: Fields with * are missing for ambiguous sequences (described in the bottom row of this table).**

Name	Type	Description
rid	int	Primary key

latin_name	string	Latin name of sequence
family	string	Specific family of sequence (e.g. 16S, 23S, 5S)
accession	string	Accession number of sequence
length	int	Length of sequence
acc_length	int	Length of the sequence in the NCBI Nucleotide database corresponding to this row's accession number. See "About accession numbers" above for more information.
seq_start	int (or NULL)	Starting index of this sequence within the NCBI sequence corresponding to this row's accession number. See "About accession numbers" above for more information.
seq_stop	int (or NULL)	Ending index of this sequence within the NCBI sequence corresponding to this row's accession number. See "About accession numbers" above for more information.
gc_content	float	Percentage (ranging from 0 to 1) of bases in sequence that are G or C
fasta_txt	str (filename)	Filename of `*.fasta` file, containing the sequence in fasta format
initial_fragment	str	The first 30 nucleotides of this sequence (in lowercase). **There is a uniqueness constraint on this column**, so MySQL will reject the addition of a new row if its value for this column already exists in the database.
orig_ct	string (filename)	Filename of `*_orig.ct` file. See "Overview" above for more information.
nop_ct	string (filename)	Filename of `*_nop.ct` file. See "Overview" above for more information.
canon_ct	string (filename)	Filename of `*_canon.ct` file. See "Overview" above for more information.
clean_ct	string (filename)	Filename of `*_clean.ct` file. See "Overview" above for more information.
*mfe_ct	string (filename)	Filename of `*_mfe.ct` file. See "Overview" above for more information.
*forced_ct	string (filename)	Filename of `*_forced.ct` file. See "Overview" above for more information.
orig_bp	int	Number of base pairs in the `*_orig.ct` structure

nop_bp	int	Number of base pairs in the `*_nop.ct` structure (excludes pseudoknots)
canon_bp	int	Number of base pairs in the `*_canon.ct` structure
clean_bp	int	Number of base pairs in the `*_clean.ct` structure
*mfe_bp	int	Number of base pairs in the `*_mfe.ct` structure
*forced_bp	int	Number of base pairs in the `*_forced.ct` structure
knots_txt	string (filename)	Filename of text file containing a list of the pseudoknots removed when creating `*_nop.ct` from `*_orig.ct`. Each base pair is listed on a new line, as the positions of the two bases separated by a space.
noncanonical_txt	string (filename)	Filename of text file containing a list of the noncanonical base pairs removed when creating `*_canon.ct` from `*_nop.ct`
isolated_txt	string (filename)	Filename of text file containing a list of the isolated base pairs removed when creating `*_clean.ct` from `*_canon.ct`
*clean_energy	float	Energy of the `*_clean.ct` structure
*mfe_energy	float	Energy of the `*_mfe.ct` structure
*forced_energy	float	Energy of the `*_forced.ct` structure
*completeness	float	Number of base pairs in `*_clean.ct` divided by the number of base pairs in `*_forced.ct`
*tp	int	The next six fields compare the base pairs in `*_mfe.ct` and `*_clean.ct`. This field is the number of base pairs that appear in both structures.
*fp	int	Number of base pairs that appear in `*_mfe.ct` but not `*_clean.ct`
*fn	int	Number of base pairs that appear in `*_clean.ct` but not `*_mfe.ct`
*precision_val (“precision” is a MySQL reserved keyword and cannot be used as a column name)	float	tp / (tp + fp)
*recall	float	tp / (tp + fn)

*f_measure	float	(2*tp) / (2*tp + fn + fp)
ambiguous	int flag (1 or 0)	Set to 1 if the sequence contains the ambiguous nucleotide symbol N
notes	str	Any notes about the sequence, to be filled in manually. This column is empty in the generated CSV.