

DNA Sequence API

A DNA sequence is a string of bases of arbitrary length. A base is represented by one of the four characters: A, C, G and T. The goal of the exercise is to define a simple API and database to store sequences using Python.

Implement methods ‘insert’, ‘get’ and ‘find’.

- The database is accessed via an instance of the SequenceDb class.
- This class offers an ‘insert’ method that has a single parameter: the DNA sequence. The DNA sequence is represented by a Python string containing any combination of the four characters ‘A’, ‘C’, ‘G’ and ‘T’. Any other character present in the string makes it an invalid input. If successful, the method stores the sequence in the database and returns a unique identifier representing the DNA string. In the event the string already exists in the database, the user should be notified. You are free to decide the specification of the unique identifier.
- The class also has a ‘get’ method that takes a DNA sequence unique identifier as an input, and returns the corresponding DNA sequence as a string. If no sequence in the database matches the identifier, the user should be notified.
- The class’ ‘find’ method takes a single parameter named ‘sample’, which is also a DNA sequence following the same rules as above.
- The find method returns a list of DNA sequence unique identifiers representing all sequences in the database that contain the sample.
- For example, if we supply the sample ‘CC’ and the database contains the following sequences:

ID	DNA Sequence
1	ACCCAGA
2	GAATAACAA
3	TCCAAT

- The return value of ‘find’ should be [1, 3] because the DNA sequences with identifiers 1 and 3 contain the ‘CC’ sequence.
- The ‘find’ method may return an empty list if no stored sequence contains the input.

Implement an ‘overlap’ method.

The SequenceDb class has an ‘overlap’ method that takes two parameters. The first parameter is named ‘sample’ and is a string representing a DNA sequence as previously described. The second is the unique identifier of a DNA sequence stored in the database, as defined

previously.

The return value of the ‘overlap’ method is a boolean. Its value is True if the DNA sequence string partially or completely overlaps with the DNA sequence with the associated unique identifier. Otherwise, False is returned. We define an overlap as a situation where a continuous series of bases are identical starting from or ending with the end of the sequence

For example, if we supply the sample as ‘GAGA’ and test it against the same database as illustrated in Level 1, we should get the following:

ID	DNA Sequence	Overlap result
1	ACCCAGA	True
2	GAATAACAA	True
3	TCCAAT	False