**Course: Advanced Bioinformatics**

**Module title: Database Searching**

**Module no. : 32**

**Database searches:** there are several objectives for database searching of genetic related data.

1. To discover or verify identity of a newly sequenced gene.
2. To find other members of a multi-gene family.
3. To classify groups of genes.

**Problems with DP approaches:**

* Dynamic programming is too slow for large databases.
* Computation and memory intensive
* Small number of sequences
* SW-algorithm with specific hardware’s

Word methods/ k-tuple methods:

* Based on heuristic methods that are not guaranteed to find an optimal alignment solution, but are significantly more efficient
* Useful in large-scale database searches.

Heuristic method

* Rules of thumb
* Not necessarily exact
* Fast and sound reasoning
* Comparing 2 sequences, exact methods such as dynamic programming are possible.
* Comparing a sequence against a database, DP consumes far too much computer time and space.

Problem: Too many calculations "wasted" by comparing regions that have nothing in common

**Initial insight:** Regions that are similar between two sequences are likely to share short stretches that are identical

**Basic Idea:** Look for similar regions only near short stretches that match exactly

**Word (k- tuple):** We define a word (or k- tuple) size that is the minimum number of exact "letter" matches that must occur before we do any further comparison or alignment.

How do we find all of the occurrences of matching words between a sequence and a database?

Could scan sequence a word at a time, but this is order L (size of database)

**Hashing:** Use a pre-computed table that lists where in the database each possible word occurs

Generation of the table is of order L (size of database) but use of the table is of order N (size of query sequence)