INTEGRATED THE DIFFERENT DATA FROM THE VARIETY DATABASE FOR QUERYING MOTIF SEQUENCE

RESEARCH MANAGEMENT INSTITUTE (RMI)
UNIVERSITI TEKNOLOGI MARA
40450 SHAH ALAM, SELANGOR
MALAYSIA

BY:

MOHD TAUFIK MISHAN
ZANARIAH IDRUS
JASMIN ILYANI AHMAD

APRIL 2014

TABLE OF CONTENT

1. Letter of Report Submission

Tarikh	:	**15 APRIL 2014**
No. Fail Projek	:	**600-UiTMKDH (PJI.5/4/1/15/12)**

Penolong Naib Canselor (Penyelidikan)
Institut Pengurusan Penyelidikan (RMI)
UiTM, Shah Alam

Tuan,

LAPORAN AKHIR PENYELIDIKAN DANA KECEMERLANGAN 'INTEGRATED THE DIFFERENT DATA FROM THE VARIETY DATABASE FOR QUERYING MOTIF SEQUENCE'.

Merujuk kepada perkara di atas, bersama-sama ini disertakan 2 (dua) naskah Laporan Akhir Penyelidikan dan satu (1) salinan *softcopy* bertajuk '**INTEGRATED THE DIFFERENT DATA FROM THE VARIETY DATABASE FOR QUERYING MOTIF SEQUENCE'** oleh kumpulan Penyelidik dari UiTM Kedah untuk makluman pihak tuan.

Sekian, terima kasih.

Yang benar,

MOHD TAUFIK MISHAN
Ketua
Projek Penyelidikan

3. Acknowledgements

We owe many people many things for the help and guidance throughout this study. The deep appreciation is extended to:

Dr. Asmadi Mohammed Ghazali
(Rektor Kampus UiTM Kedah)

Dr. Mahadir Ismail
(Timbalan Rektor Penyelidikan & Jaringan Industri, UiTM Kedah)

and never forgotten

Special thanks to PJI UiTM Kedah for the cooperations and supports us for carrying out this study.

Thank you.

5.2 Enhanced Executive Summary

In performing protein secondary structure prediction procedures, biologists need to use variety types of sequence data from multiple biological repositories which are available publicly in the Internet. A lot of researches have been done in minimizing the numbers of repositories needed for the prediction procedures. However, due to the size complexity and numbers of repositories used has created a major challenge in integrating all different data into one repository or database. This challenge is known as syntactic heterogeneity problem. The aim of this research is to overcome the problem by transforming all the different data form variety of databases such as Prosite, Blast, Print and PDB into flat file format and other format into relational form using XML and asp dot net. From studies that have been conducted, XML approach is considered as a better choice for biological data integration. And this research has reveals that query made from relational database incorporating XML schema gives better query performance after integrating the variety data into one repository or relational database using metadata framework. As a result, this research showed some tool can search different data and different sizes of protein secondary structure data stored in the relational database and the result can be retrieved faster and reliable.