Honours Project Proposal - Species Identification through DNA String Analysis

Mark Vorster
Supervisor: Prof. Philip Machanick

28 February 2012

1 Principal Investigator

Mark Vorster
12 Croft Street
Grahamstown, 6139
072 618 7960
g07v3343@campus.ru.ac.za
Supervisor: Professor Philip Machanick

2 Project Title

The provisional title of the project is: *Species Identification through DNA String Analysis*

3 Statement of the Problem

The Rhodes University Department of Biochemistry, Microbiology & Biotechnology has found need to identify the distinct species of bacteria given a large set of DNA sequences, however without existing tools to solve this specific problem they have found it takes an inordinate amount of time. The objectives of the research is to explore bioinformatic and data mining techniques and algorithms in order to develop a tool to allow them to solve the problem quickly and conveniently.

4 Background

In collaboration with Rhodes University Department of Biochemistry, Microbiology & Biotechnology who, in their bacteria research have found need to identify a number of different bacterial species in an aquatic sample, have found

the existing software tools lacking. In an interview with Professor Dorington she described some of the problems that they are having which include analysing genes of various lengths and analysing large sets of data, upward of ten thousand samples. She explained that they could work with small samples but that it did not scale to the proportion they require, and that the small samples only yielded usable data for the most predominant of the bacterial species, whereas they are most interested in the less common specimens. Currently they are working with sections of the genes that are about 400 bases long (but hope to improve their techniques to be working with sections up to 800 bases) and that in the section a difference of four or less of those bases (1%), with other constraints and considerations due to the difficulties experienced with reading the gene, would indicate the same species. The existing tools for this task have been found to take up to ten days to analyse their samples, but, given that the sequences can be preprocessed to isolate this gene of interest such that they can be assumed to be globally aligned. With this knowledge and the fact that a specific tool would not require to calculate the result of an alignment and could therefore prune branches that cannot yield a result quickly, it should be possible greatly reduce the amount of time it takes to analyse their data sets.

5 Approach

The development of this project will undergo multiple steps.The project will begin with a review of bioinformatic and string analysis literature. Ideally the first step would be to analyse the current system's algorithms, but as the current system is a commercial product the source code is likely to be unavailable. This will be followed by the design of a system that the bioinformaticians would be able to utilise in their research, this would require further meetings with them to discuss their requirements. The system can then undergo an iterative implementation with testing and further design. The program will need to be performance tested to ensure results are accurate and reliable.

6 Proposed Timeline

2nd March	Formal Written Proposal
6th March	Initial Seminar to Department
12th March	20 Papers Found (broad goal)
5th April	Investigation into Existing Tools
20th April	Implementation Plan in Place
30th April	Draft Literature Review
28th May	Final Literature Review
24th July	Seminar 2
31st August	Implementation Completed
29th October	Seminar 3
2nd November	Thesis Completed
5th November	Research Website Complete
21st November	Final Research Oral Examination