THE STEINER PROBLEM IN GRAPHS

AND ITS APPLICATION TO PHYLOGENY

A dissertation presented by

M. L. Shore

in partial fulfilment of the requirements

for the degree of

Master of Arts in Computer Science

at

Massey University

November 1979

ABSTRACT

In this thesis we consider two problems, one in Graphy Theory and the second in Evolutionary Biology. The first problem, the Steiner Problem in Graphs, belongs to the class of problems known as NP-complete. That it belongs to this class is a measure of its difficulty; if a polynomial solution can be found for the Steiner Problem in Graphs, then by definition a polynomial solution will have been found for all other NP problems. For this problem we present a solution method based upon a branch and bound approach, and we show that it complements the current methods available for solving this problem, allowing solutions to the Steiner Problem in Graphs to be calculated for any graph of size $\leq$ 30 points in reasonable computing time.

The second problem is associated with evolution, and is an application of the Steiner Problem in Graphs. In trying to determine the evolutionary path from some common ancestor to existing species, a tree may be drawn to show these paths. This tree is called a phylogenetic tree, or phylogeny. There must be some criterion for deciding which of the many phylogenies that may be drawn most closely resembles the actual evolutionary changes. The criterion used in this thesis, used by many researchers in this area, is that of minimising the total number of changes in the phylogeny. It is shown that this problem is similar to the Steiner Problem in Graphs, and various solution methods based on heuristic graph theoretical techniques are discussed. Subsequent to this, a method of proving a phylogeny optimal is looked at, and an extension to this method presented which will allow larger phylogenies to be analysed.

ACKNOWLEDGEMENTS

The direct and indirect encouragement and support of many people has been greatly appreciated during the preparation of this thesis. My special thanks go to Drs Foulds, Gibbons, Hendy and Penny without whom this thesis would not have been possible.

Thanks also go to the Ministry of Defence EDP Directorate for their direct and indirect support throughout the duration of this thesis.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 The scope of this thesis

As its name suggests, this thesis examines two problems, one in graph theory and the other in evolutionary biology.

The Steiner Problem in Graphs (SPG) is a lesser known problem of combinatorial optimization, and although methods for its solution do exist, the size of the problem that can be solved still remains severely limited. The problem belongs to that class of problems known as NP-complete; by this we mean that the problem could be solved in polynomial time on a non-deterministic machine, and all other NP problems can be reduced to it in polynomial time. Thus if any NP-complete problem can be solved with a polynomial algorithm, then all NP problems are polynomially solvable, hence this class of problems are very difficult to solve. The thesis discusses two current methods of solving the problem - one by Hakimi [1971] and the other by Dreyfus and Wagner [1972]. There exists another solution method, that of Levin [1971], but this paper was discovered too late to be included in this thesis. However, from the expressions given for algorithm solution time, the method of Levin will, for non-trivial problems, always take longer than the method of Dreyfus and Wagner. The aim of this part of the thesis is to explain the techniques used in solving the SPG, in order to attempt similar methods to solve the phylogeny problem. At the same time, it is aimed to present a new algorithm for the SPG, and to show that this approach compares favourably with existing solution methods.

The concept of phylogeny - the building of a tree in which points represent species and lines represent evolutionary changes between species - is discussed

in the second part of this thesis. The various methods of building phylogentic trees are discussed briefly, and then a recent method, due to Foulds, et. al. [1978] of generating phylogenies using graph theoretical techniques is looked at in depth, and attempts made to use this and other similar techniques to allow completely automatic generation of trees. It is not known, at this time, whether the phylogeny problem is NP-complete. Then the latest approach used by Hendy et. al. [1979] is examined. This method involves two phases; the building of the tree, and proving that the tree is optimal. Apart from the automatic generation of trees, this part of the thesis aims to present an extension to the proof technique. Due to time constraints, this new approach was not able to be fully implemented and evaluated for inclusion in this thesis, but preliminary results are discussed.

Although a great deal of computer programming was involved in the preparation of this thesis, in evaluating and comparing algorithms for both the SPG and the phylogeny, the only program listing included as part of this thesis is the SPG branch and bound solution method algorithm program, and that is given as an appendix.