NOVEL MATHEMATICAL MODELS FOR BIOLOGICAL DATA

By

Phaedra Agius

A Thesis Submitted to the Graduate
Faculty of Rensselaer Polytechnic Institute
in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY
Major Subject: Mathematics

Approved by the
Examining Committee:

_____
Professor Kristin Bennett, Thesis Adviser

_____
Professor John Mitchell, Member

_____
Professor George Plopper, Member

_____
Professor Michael Zuker, Member

Rensselaer Polytechnic Institute
Troy, New York

May 2007
(For Graduation August 2007)

NOVEL MATHEMATICAL MODELS FOR BIOLOGICAL DATA

By

Phaedra Agius

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Mathematics

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Professor Kristin Bennett, Thesis Adviser
Professor John Mitchell, Member
Professor George Plopper, Member
Professor Michael Zuker, Member

Rensselaer Polytechnic Institute
Troy, New York

May 2007
(For Graduation August 2007)

CONTENTS

ABSTRACT

The interdisciplinary field of mathematical biology aims to model biological processes using mathematical tools and techniques. This work presents the computational and statistical processes involved in the interpretation and modeling of three types of biological data.

Our first project involves the analysis of a set of preprocessed DNA sequences of the *spa* gene in the bacterium *Staphylococcus aureus*. We develop hybrid sequence algorithms to compare the sequences and, using the resulting comparative scores, we use a clustering algorithm to group the bacteria into family types. We validate our results by contrasting our family types to those pre-established by existing techniques. Our methods are cheaper than existing methods, can replicate the known labels and may be fine-tuned to be more discriminative.

In our second study we present a novel algorithm, the Relaxed Base Pair (RBP) score, that compares RNA secondary structures in a more relaxed way than the Base Pair (BP) metric. This method has a tunable parameter t which, when driven to zero, transforms it into the Base Pair metric. Our results indicate that RBP scores induce more consistent and significant clusterings than the BP metric. Using the multiple clusterings obtained with various t values, one may define multiple cluster centroids. These may be clustered to reveal common underlying structures within a sample of RNA secondary structures.

Often biological experiments are performed multiple times, so that one experiment is represented by a set of results, or replicates. In our third project we develop regression models for replicate data. Existing regression methods can only regress to one-to-one data, forcing such methods to regress to all of the replicate data or to the replicate means. Our new regression model, the ReV machine, regresses to the replicate ranges. Our results indicate this new method to perform as well as existing methods, even outperforming them for certain types of data.