ALGORITHMS AND SOFTWARE FOR NUCLEIC ACID SEQUENCES

By

Nicholas R. Markham

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: Computer Science

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

Examining Committee:

Michael Zuker, Thesis Adviser
Christopher Bystroff, Member
Mukkai S. Krishnamoorthy, Member
Chjan C. Lim, Member
Malik Magdon-Ismail, Member
Lee Newberg, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2006
(For Graduation May 2006)

Abstract

We present here the design and implementation of a general model for ensembles of one or two nucleic acid sequences. Unlike the traditional two-state model, our model does not require that strands be complementary or nearly so; nor does it assume equal strand concentrations. Furthermore, our model considers the full ensemble of possible states — foldings, homodimers and heterodimers.

The core of the model comprises several algorithms for computing both minimum energies and partition functions for foldings and for hybridizations with and without intramolecular base pairs. Most of these algorithms have been described previously, but we present them in a unified form for the first time. An important tenet of our model is that a single-stranded base may, but need not, dangle on an adjacent base pair. Unlike the Vienna RNA Package, which assumes that all bases adjacent to a closing pair of a helix participate in single-base stacking, our UNAFold software considers (in general) four states for each motif: no base dangling, the 5' base dangling, the 3' base dangling and both bases dangling.

Though these algorithms are valuable in their own right, their results can also serve as input for additional algorithms that simulate characteristics of ensembles. In particular, we describe the computation of mole fractions, thermodynamic quantities and UV absorbance for one- and two-sequence ensembles. We also describe a method to obtain the enthalpy of each species, and therefore that of the ensemble, using stochastic tracebacks.

All of these algorithms have been implemented in the UNAFold package. This robust, portable software is composed of C programs and Perl scripts that facilitate the simulation of an ensemble as well as allowing for numerous other, auxiliary calculations. UNAFold is well suited both for comparing theoretical simulations to experimental results and for predicting quantities not measurable directly.

UNAFold also implements a model for energy in unfolded single strands, which has greatly improved predictions of enthalpy, entropy and melting temperature for short duplexes.