



Inverse RNA Folding

E.Sizikova, T. Hyland, A.Badugu, J. Anderson, R. Lyngsø and J. Hein[‡]
Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG, United Kingdom
[‡] hein@stats.ox.ac.uk

Motivation

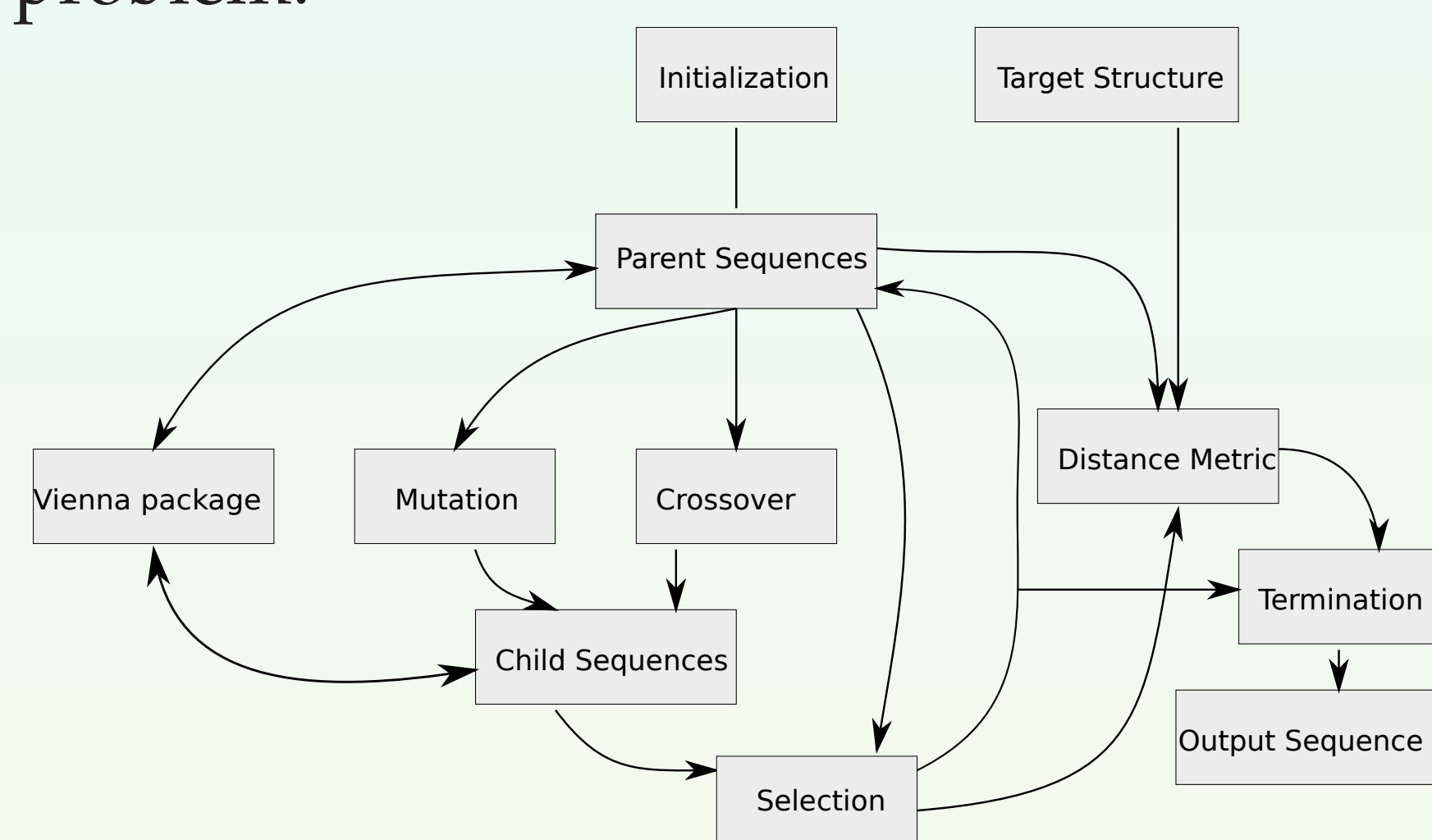
RNA molecules, the fundamental building blocks of life, are chains of nucleotides that can bond to parts of themselves. This allows them to form complex structural motifs that can be used as building blocks for self assembling nano-scale machines.

One of bioinformatics' challenges is to design an implementation that, given an input structure, would predict a sequence that would most likely fold into this structure at a given temperature. This is known as the **Inverse RNA folding problem** for single structures. We have developed a genetic algorithm based solution to this problem.

Some RNA molecules have the property of folding into different structures at different temperatures. This can be exploited in applications for synthetic biology. For example, one can design a simple latching mechanism with an RNA stand at the binding site where simple changes in temperature can allow/block binding. We extended our software to predict a sequence which folds into several different structures at differing temperatures.

Flow of the Algorithm

Genetic algorithms are part of a wider class of evolutionary algorithms which mimic natural evolution to obtain solutions to an optimization problem.



Genetic Algorithms are usually split into several sub-parts: initialisation, selection, reproduction and termination. Above figure illustrates how our genetic algorithm solves the Inverse RNA Problem.

Initialization

In the initialisation phase an initial population of possible solutions are generated. These may be generated completely randomly.

Selection

The selection phase involves picking solutions that will become parents for the next generation of the algorithm. We would like to generate better solutions to the optimization problem with each subsequent generation, so we try to pick parents from those solutions that come closest to solving the problem.

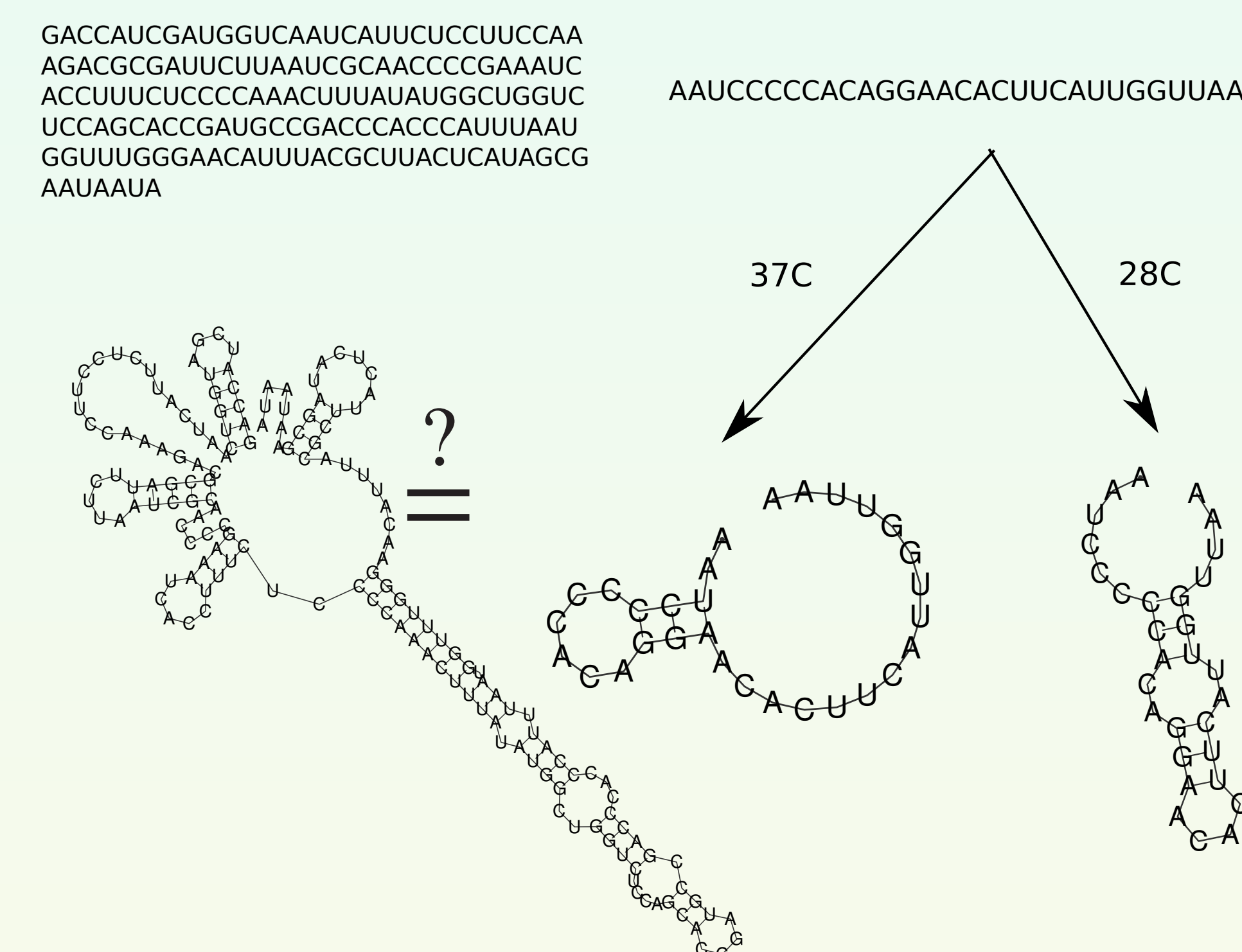
Reproduction

The reproduction phase involves generating child solutions from the parents we picked in the selection phase. The two most common types of reproduction are **mutation** and **crossover**. Mutation takes a single parent and changes one or more positions in its chain to create a child solution. Crossover takes two parents' chains, and cuts them up into pieces, creating two child solutions by sticking together these pieces in an appropriate order.

Termination

The algorithm is terminated after either, an exact solution has been found or a predefined maximum number of generations has been reached.

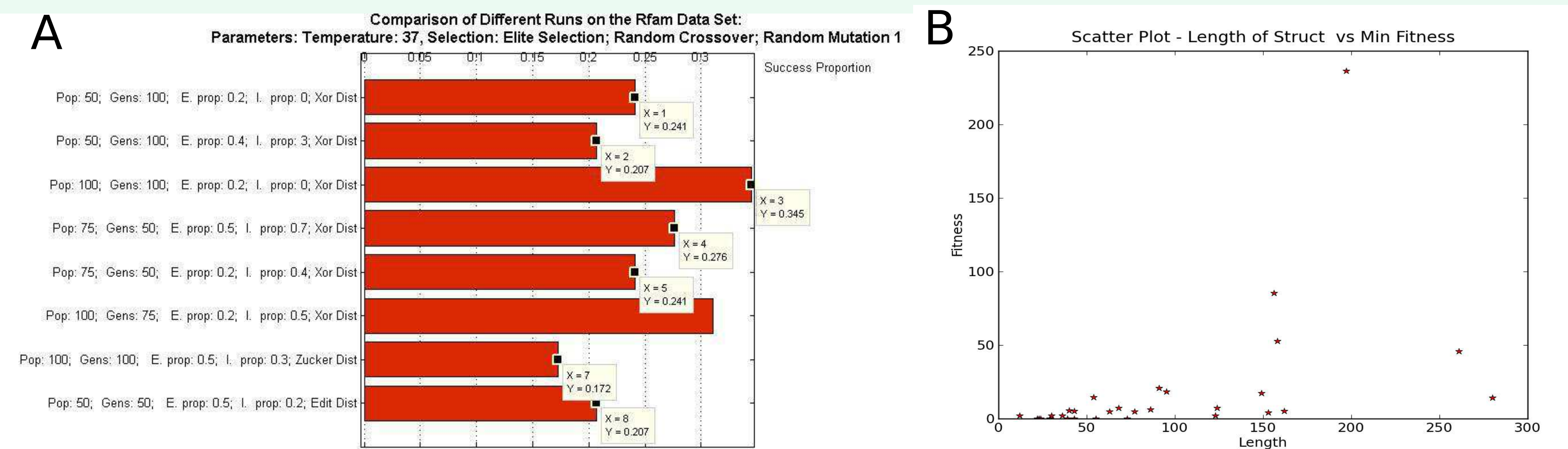
Multiple Target Extension



Single Target Results

We used two different data sets to test our algorithm. The first was a sample data set generated from known sequences using Vienna Fold. The second was a list of 29 real RNA sequences called the Rfam data set. Using the generated data we successfully predicted 88.7% of the structures. Whereas for the Rfam data set, we correctly generated 34.5% of the sequences.

Data Analysis



Extensions and Improvements

We believe that we could definitely improve our algorithms and increase the current success rates by a reasonable amount. For the Single Target algorithm we would want to get a success rate of over 51% on the real data set, in order to be up to speed with other available implementations.

There is currently no available solutions to the Multiple Target Problem (dealing with riboswitches). We now offer a solution with a success rate of 33%, but would like to further improve this result, possibly by considering previously avoided computation parameters, such as MFE energies and other distance metrics.

References

- [1] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multistable RNA molecules. *RNA*, 7(2):254-265, 2001.
- [2] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105-1119, 1990.
- [3] A. Taneda. MODENA: a multi-objective RNA inverse folding. *Advances and Applications in Bioinformatics and Chemistry*, 4:1-12, 2011.
- [4] R. Lyngsø. Lecture notes. RNA secondary structures, 2010.
- [5] T. Cormen, C. Leiserson, R. Rivest, C. Stein. *Introduction to Algorithms - Third Edition*. The MIT Press, 2009.
- [6] M. Zucker. On Finding All Suboptimal Foldings of An RNA Molecule. *Science*, New Series 244:4900, 1989.

Multiple Target Results

For the Multi Target algorithm we used a data set provided by the Regulatory RNA group. These RNA molecules had been folded at 29 and 37 degrees by UNAFold, a different folding implementation. We found it to be more difficult to predict solutions to this dataset, possibly because UNAFold and Vienna may fold sequences differently. Out of the 30 sequences we managed to get exact matches to 10. However, even in the cases when we did not get an exact match we still generated solutions that had very similar structures to both targets.

Acknowledgements

We would like to thank our supervisors, Rune Lyngsø, James Anderson, Jotun Hein, as well as Adam Novak and Joe Herman, for continuously providing advice and guidance in the course of our project. This work was carried out as part of the Oxford Summer School in Computational Biology, 2011, in conjunction with the Department of Plant Sciences and the Department of Zoology. Funding was provided by the EU COGANGS Grant. We thank Dr. Steven Kelly for providing computational resources.