Final Project Proposal

Robert Toscano and Shannon Iyo

November 5, 2007

1 Overview

For our final project, we propose to combine the brute-force computational capabilities of computer algorithms with the intuitive and creative reasoning abilities of humans to increase the accuracy of gene sequence alignment. This will involve the creation of a graphical user interface that will allow a human to intervene during the sequence alignment algorithm. For example, as the human watches the computer algorithm perform the alignment, the human might recognize (using his/her intuition) that a particular high scoring local alignment is incorrect. Not only could the human correct the local alignment in real-time, but the algorithm would recognize this as an invariant constraint and continue its algorithm appropriately. As a consequence, the algorithm will use this human intervention event to learn (update its scoring matrices appropriately).

2 Constraint Invariants

While the sequence alignment algorithm is running, the user will have the ability to interrupt it and assign a constraint on the alignment. A constraint on the sequence alignment would mean that no matter how the algorithm aligns the two sequences, some substring of the two sequences will have a guaranteed alignment. For example, as the user watches the alignment take place, and he/she realizes that the alignment in some substring of the two sequences is off by one nucleotide, the user can correct the alignment by sliding one of the substrings to the left or right and locking that substring alignment in place. When the algorithm continues to run with the constraint in place, that human assigned substring alignment will stay invariant. Eventually, the alignment will contain islands of human issued substring alignments surrounded by computer generated alignment between.

3 A Modified Sequence Alignment Algorithm

We propose to build our application upon the Needleman-Wunsch global alignment algorithm. A more advanced alignment algorithm such as LAGAN [1] would be preferable but the implementation of such an algorithm is probably outside the scope of this project.

Allowing the user of this application watch the algorithm as it works is crucial in this interaction. Our sequence alignment algorithm will be almost identical to the Needleman-Wunsch algorithm except in the following aspects:

- perceived running time of the algorithm
- trace back direction
- the scoring matrix can be updated in real time

The perceived running time of the algorithm is the actual running time of the algorithm plus additional throttling due to the user interface. Because we want the algorithm to be able to be viewed in realtime, the trace back phase of the Needleman-Wunsch will be throttled down for appropriate user perception.

If the user creates a constraint invariant, the alignment algorithm should propagate outwards from the point of the constraint to the left and right. This way, the user can see the computer's changes being made to the alignment in the near area of the constraint. This helps with the usability of the interface.

Our application will incorporate the user's input (input constraint invariants) into making the alignment algorithm smarter. The algorithm's scoring matrix will be updated according to the substring alignment enforced by the user's constraints. This is to say that a user's intervention can only make the alignment better than whatever the computer can produce. We plan to implement this part of our application using a neural network or another simple machine learning algorithm.

4 User Interface

We would like to present the user with an efficient and rich interface for exploring sequence alignments. As a preliminary design, the interface will have a large area dedicated to viewing and manipulating alignments between two strands of DNA. The interface will provide a means for zooming in and out, and allow the user to tag alignment regions as correct or incorrect, as well as manually creating new alignment regions. It will likely be useful to show information regarding sequence identity in the alignments, which could be done using color-coding (e.g. high-identity regions in blue, low-identity regions in red).

5 Evaluation

To evaluate the performance of our algorithm in terms of accuracy of sequence alignment, we will compare alignments generated by our human-intervention capable interface with (1) alignments based on the same alignment algorithm but no human interaction, (2) alignments generated by other algorithms, and/or (3) known or widely accepted sequence alignments. The human intuition required for effective human intervention can be supplied by a professor or researcher that has had experience with aligning sequences.

6 Contributions

This project will demonstrate a method for (a) placing a human expert in the critical path of sequence alignment and (b) learning alignment scores from human expertise. We will consider the project a success if we can demonstrate an increase in alignment performance compared to a machine-only alignment process.

7 Task summary

- Implement alignment algorithm
- Design and implement user interface
- Implement learning and score updating (from user interaction)
- Evaluate system (requires human expert)

References

[1] Michael Brudno, Chuong B. Do, Gregory M. Cooper, Michael F. Kim, Eugene Davydo v, NISC Comparative Sequencing Program, Eric D. Green, Arend Sidow, and Serafim Batzoglou. LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13:721–731, 2003.