Segment Match Refinement and Applications

Aaron L. Halpern,
Daniel H. Huson*
& Knut Reinert

CELERA

* Now at: Center for Bioinformatics Tübingen

Overview

- Motivation: comparison of HGP and Celera assemblies of human
- Problem: one-to-one matching of sequences
- Solution 1: A greedy heuristic
- Solution 2: Optimal refinement algorithm
- Other applications
- Comparison of solutions

Two assemblies of human

Comparison of two assemblies

- What do they have in common?
  - Compute one-to-one matching of sequence
- How do they differ?
  - Determine how much of the remaining difference is unique,
  - unique-repetitive or
  - unique-under-collapsed sequence

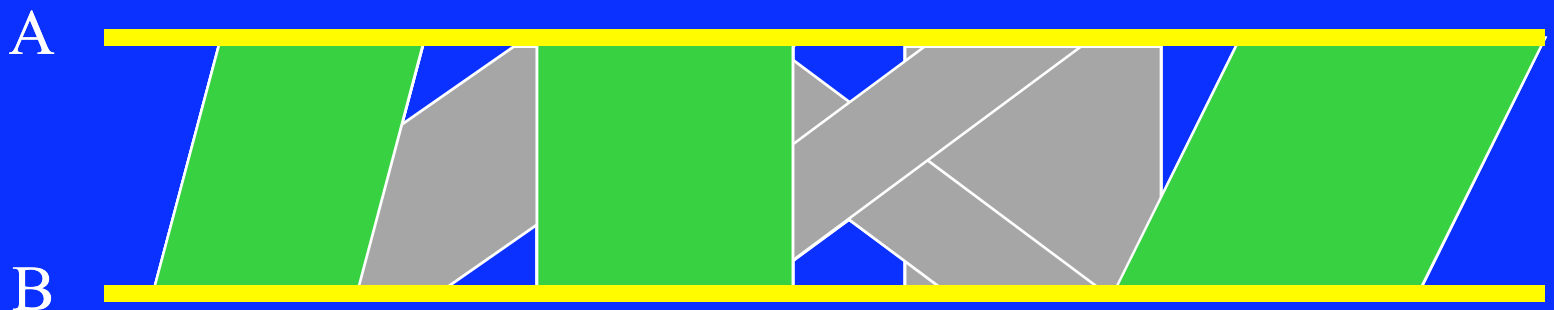
Comparison of HGP and Celera assemblies

The Matched Sequence Problem

Given sequences A and B and a set Σ of segment matches between them. The Matched Sequence Problem is to compute a set of non-intersecting matches Σ' that are all submatches of Σ, such that the amount of sequence covered by the matched segments is maximized

Greedy one-to-one matching

- Given a set Σ of segment matches between two sequences A and B:

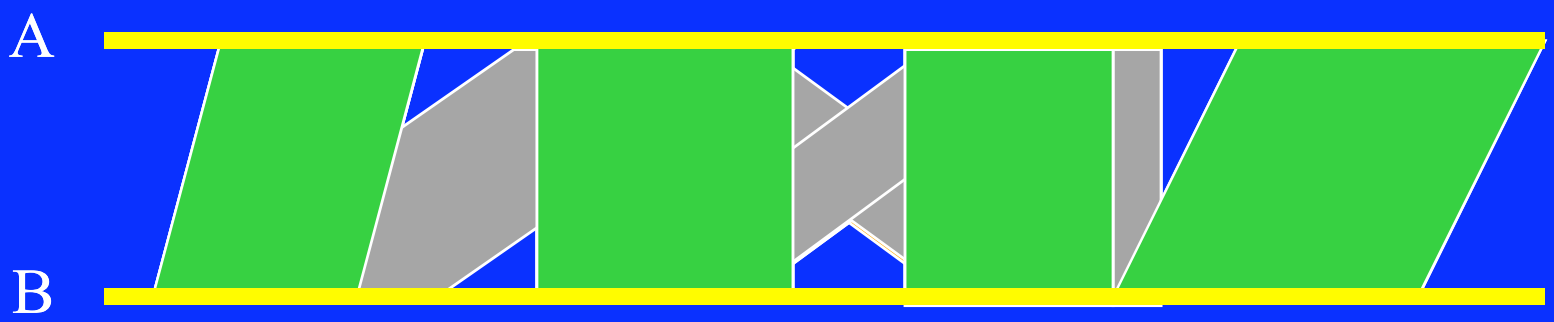


- A one-to-one matching of sequence can be obtained by greedily selecting a set of disjoint matches using a priority queue

Daniel Huson, 2002

Greedy matching with trimming

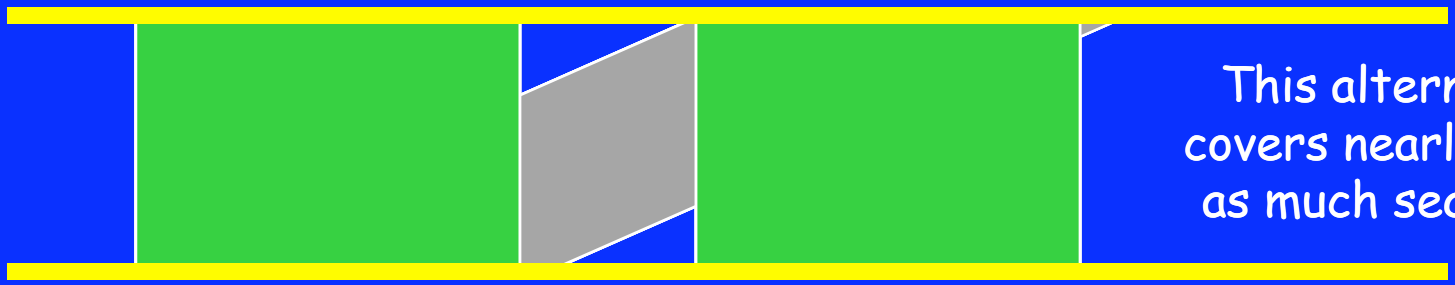
- Given a set Σ of segment matches between two sequences A and B:



- Trim matches that partially overlap with already selected ones and re-insert into priority queue

How well does greedy selection with trimming work?

- Goal: obtain a set of non-overlapping matches that cover as much of both sequences as possible
- Greedy selection runs fast and obtains results that seem reasonable…
- BUT it could missing up to half of the attainable coverage:



Daniel Huson, 2002

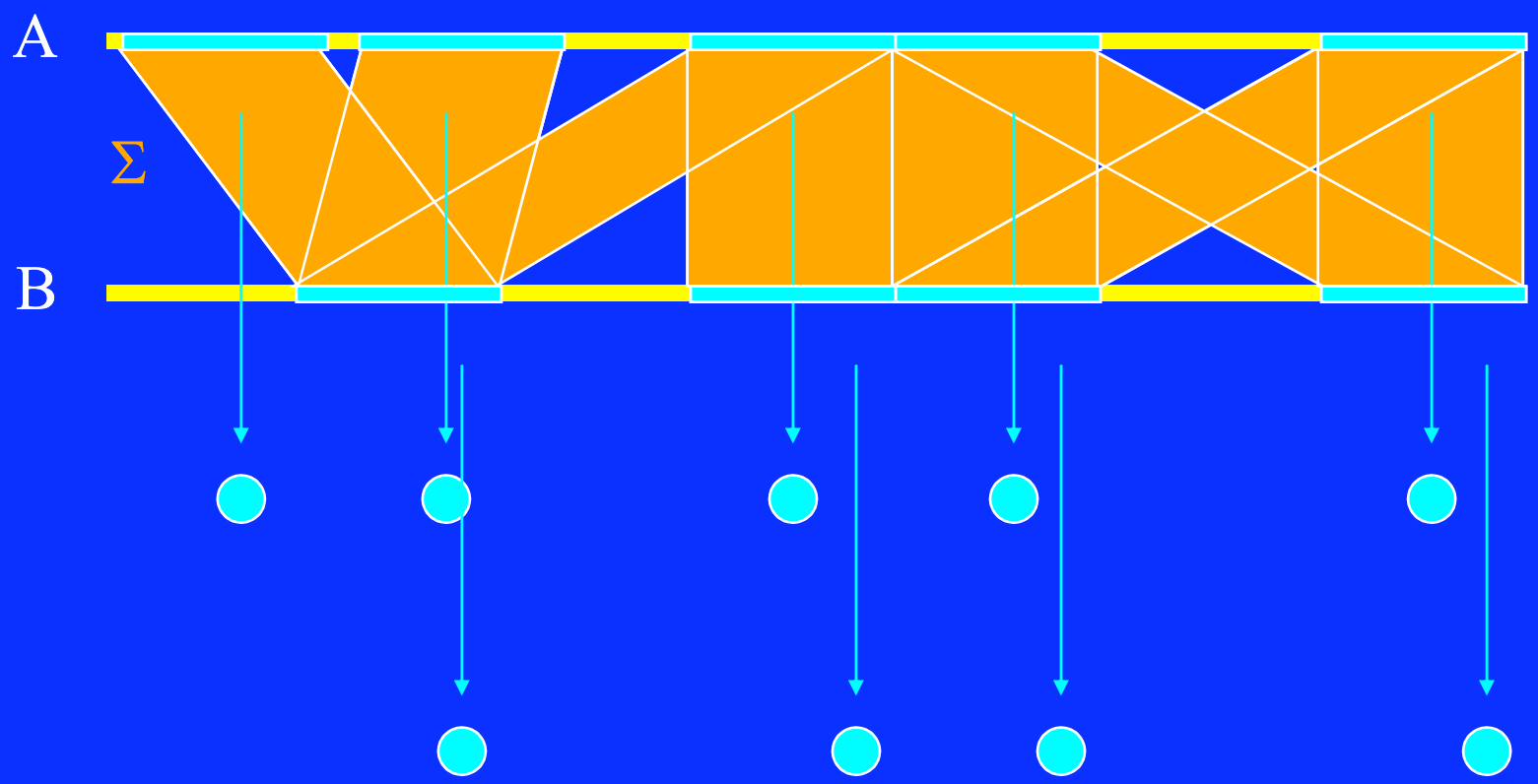
How about an optimal matching?

- Can an optimal set of non-overlapping matches be computed efficiently?
- How much better is an optimal solution than the one obtained by the greedy+trimming approach?

Sequence matching vs. graph matchings

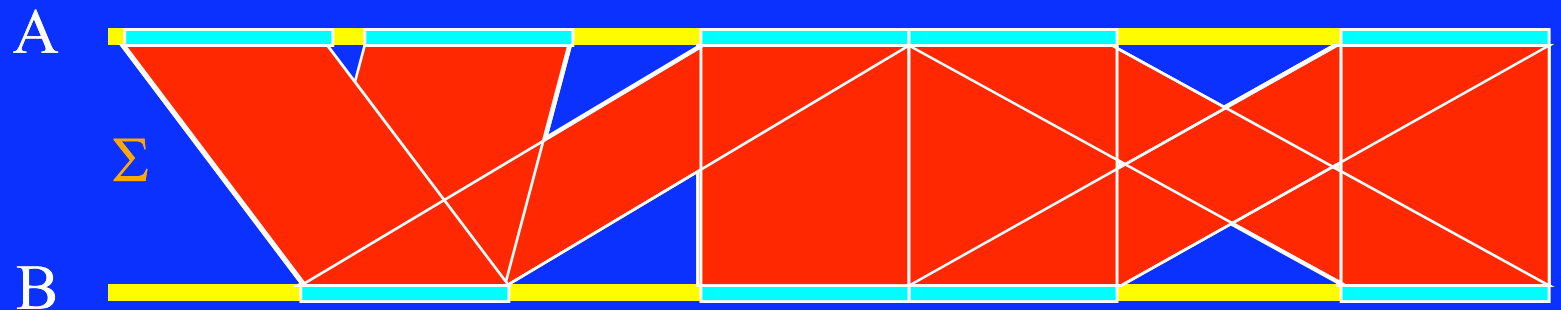
- Given two sequences A and B and a set Σ of segment matches
- Assume that Σ is *resolved*, i.e. that the projections of any two matches onto each sequence are either disjoint or identical
- Then, the Sequence Matching Problem can be reduced to a weighted matching problem for a bipartite graph

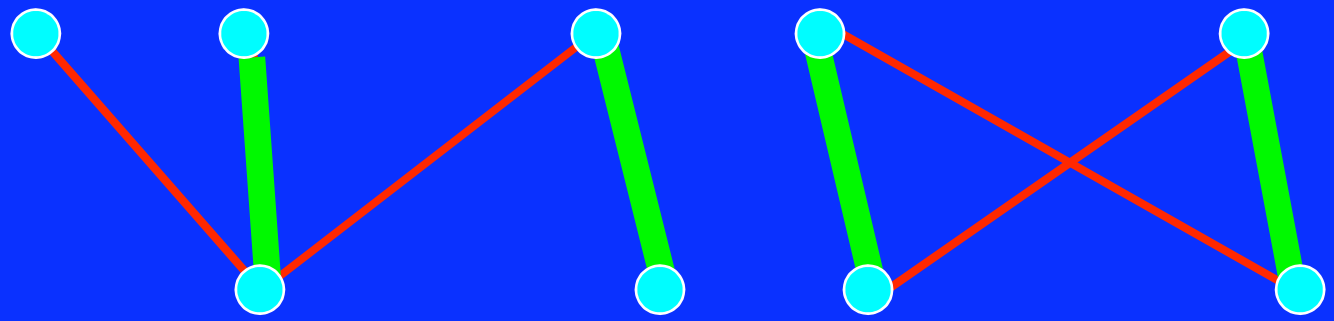
The match graph



Daniel Huson, 2002

The match graph





Heaviest matching in bipartite graph gives optimal sequence matching

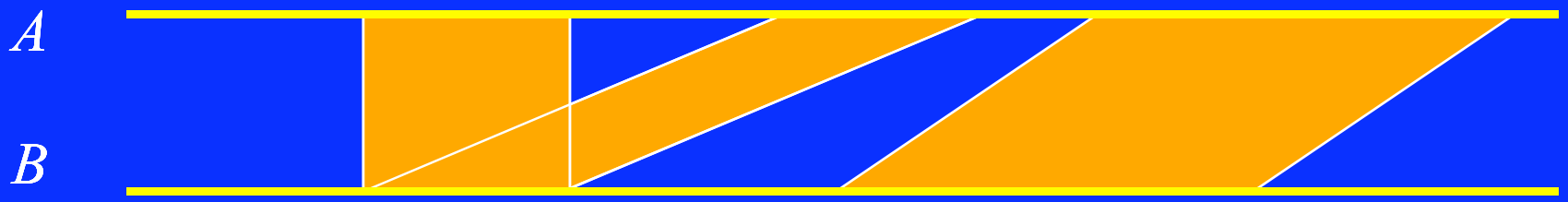
Daniel Huson, 2002

The resolved refinement

- Given a set Σ of matches between A and B.
- To be able to apply the bipartite graph maximal matching algorithm, we need to refine all overlapping matches so as to obtain a resolved refinement Σ' of Σ

Resolved segment matches

- Given a set Σ of *segment matches* between two sequences A and B



- Want to determine a *resolved* subset Σ' of Σ in which the projections of any two matches onto each sequence are either disjoint or identical.

Daniel Huson, 2002

In other words…

- Given many overlapping segment matches produced by programs such as MUMer or BLAST
- Determine a resolved set of matches that do not overlap inproperly