Introduction to Assembly Algorithms

Mihai Pop

Recap...

- Sequencing technologies only "read" small chunks of DNA, yet genomes are substantially larger
- The shotgun sequencing approach generates many random fragments from the original DNA
- The task of the assembly program is to stitch together the many small pieces into a reconstruction of the genome

- Essentially..... a huge jigsaw puzzle

- Think: shred a collection of Harry Potter books at random then try to rebuild the original without any additional information.

Assembling two cities

it was the best

was the *age of*

best *of times* **it**

it was the age *of times* **it was**

wisdom **it was the**

it was the best

was the best *of*

the worst *of times*

was the worst *of*

was the best *of*

times **it was the**

it was the *age*

times **it was the**

was the *age of*

the best *of times*

worst *of times* **it**

age of wisdom **it**

it was the *age*

of wisdom **it was**

it was the worst

the *age of* wisdom

of times **it was**

the *age of* foolishness

Shortest common superstring problem

What are we looking for? (mathematically)

Given a set of strings, $\Sigma=(s_1, ..., s_n)$, determine the shortest string S such that every s_i is a sub-string of S.

NP-hard

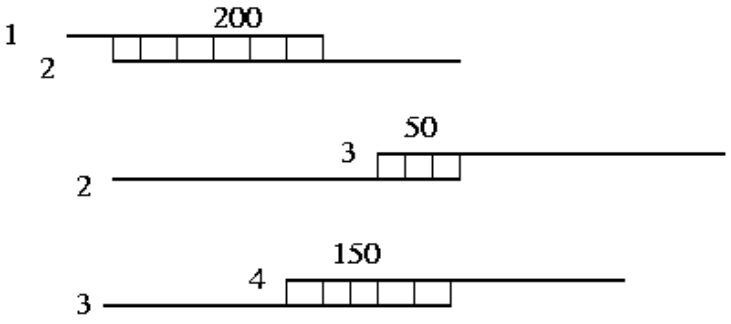
approximations: 4, 3, 2.89, ...

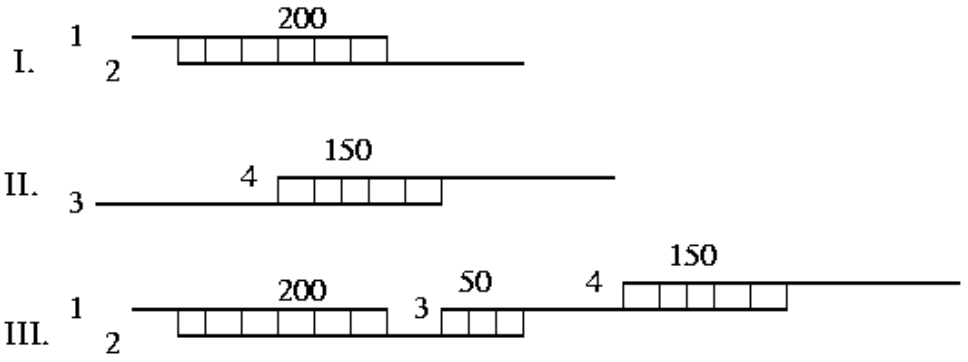
```
...ACAGGACTGCACAGATTGATAG
      ACTGCACAGATTGATAGCTGA...
```

Greedy algorithm details

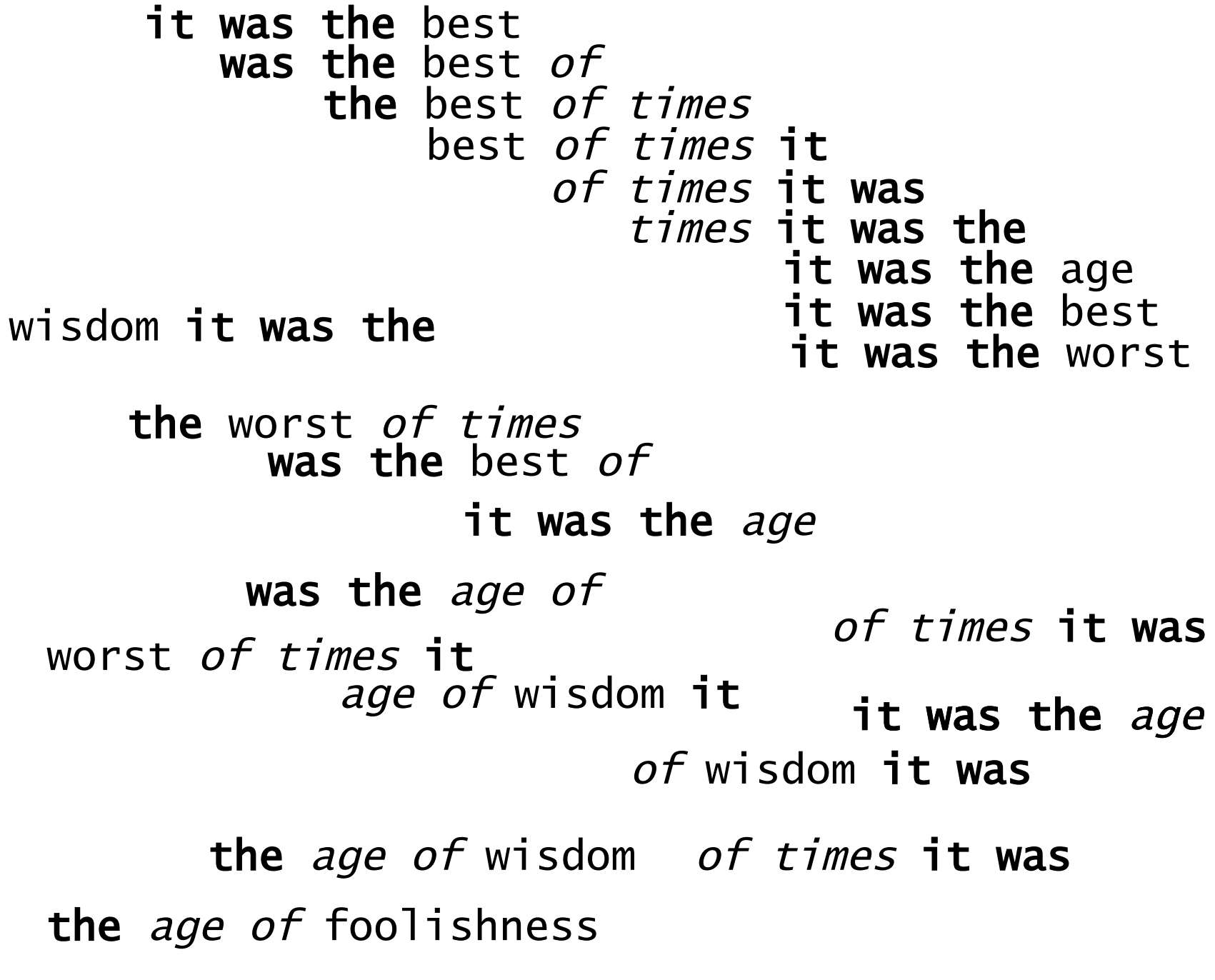
- Compute all pairwise overlaps
- Pick best (e.g. in terms of alignment score) overlap
- Join corresponding reads
- Repeat from * until no more joins possible

Greedy algorithm (4-approximation)





Greedy approach gets 'stuck'

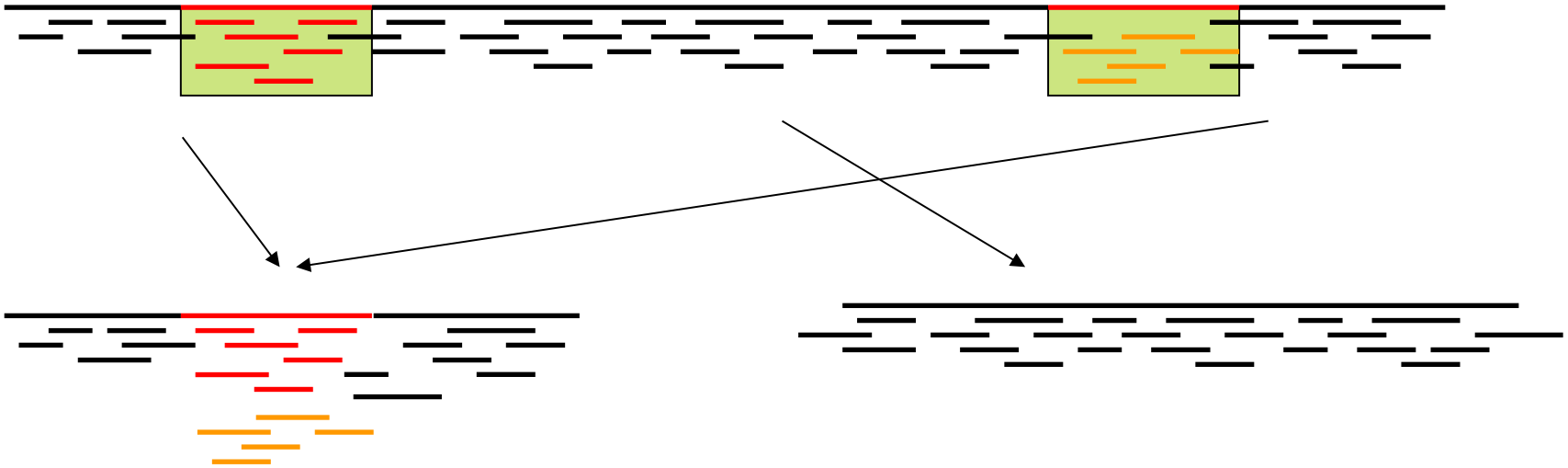


Repeats (where greedy fails)

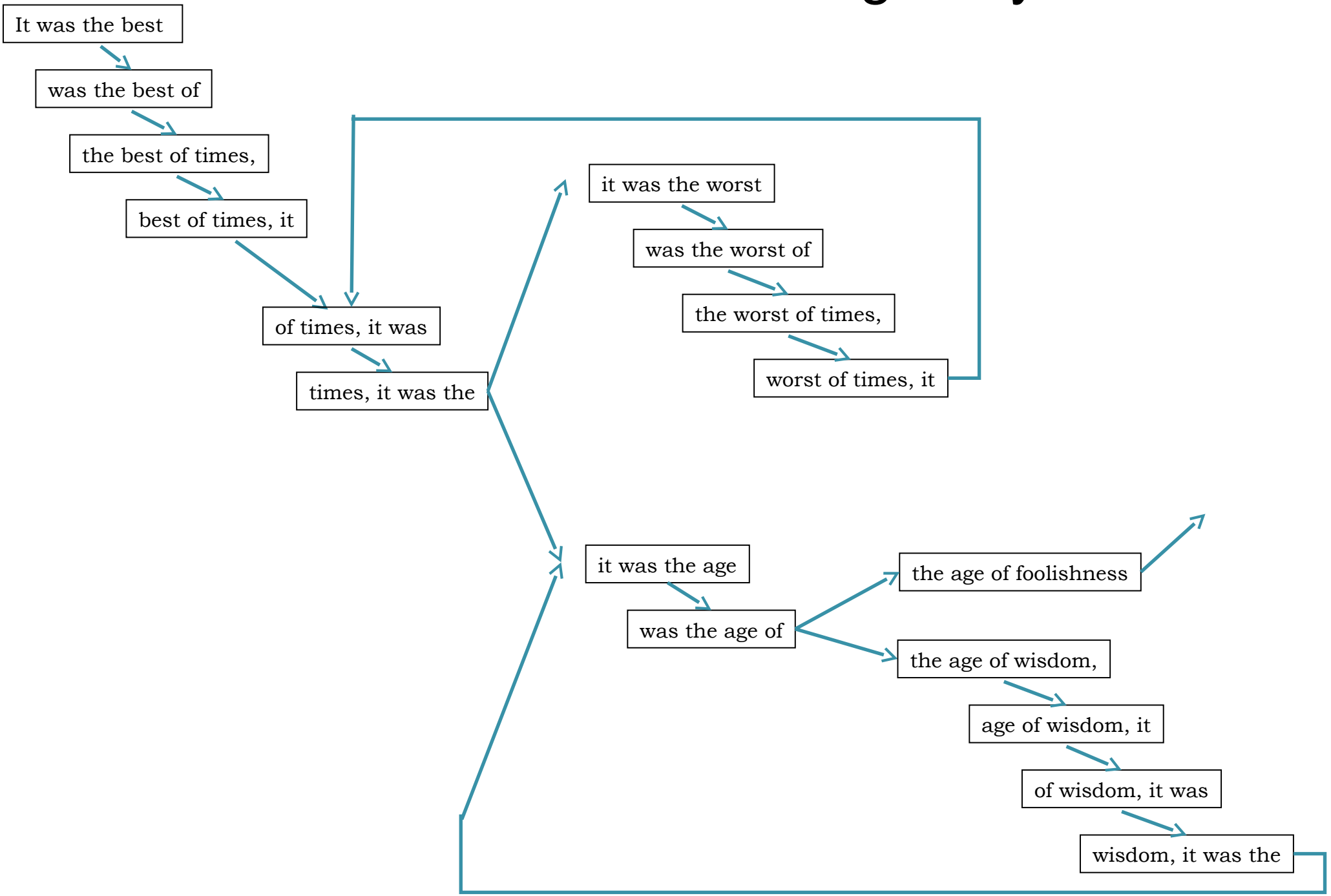
AAAAAAAAAAAAAAAAAAAA

AAAAAA AAAAAA AAAAAA
AAAAAA AAAAAA
AAAAAA AAAAAA

AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA



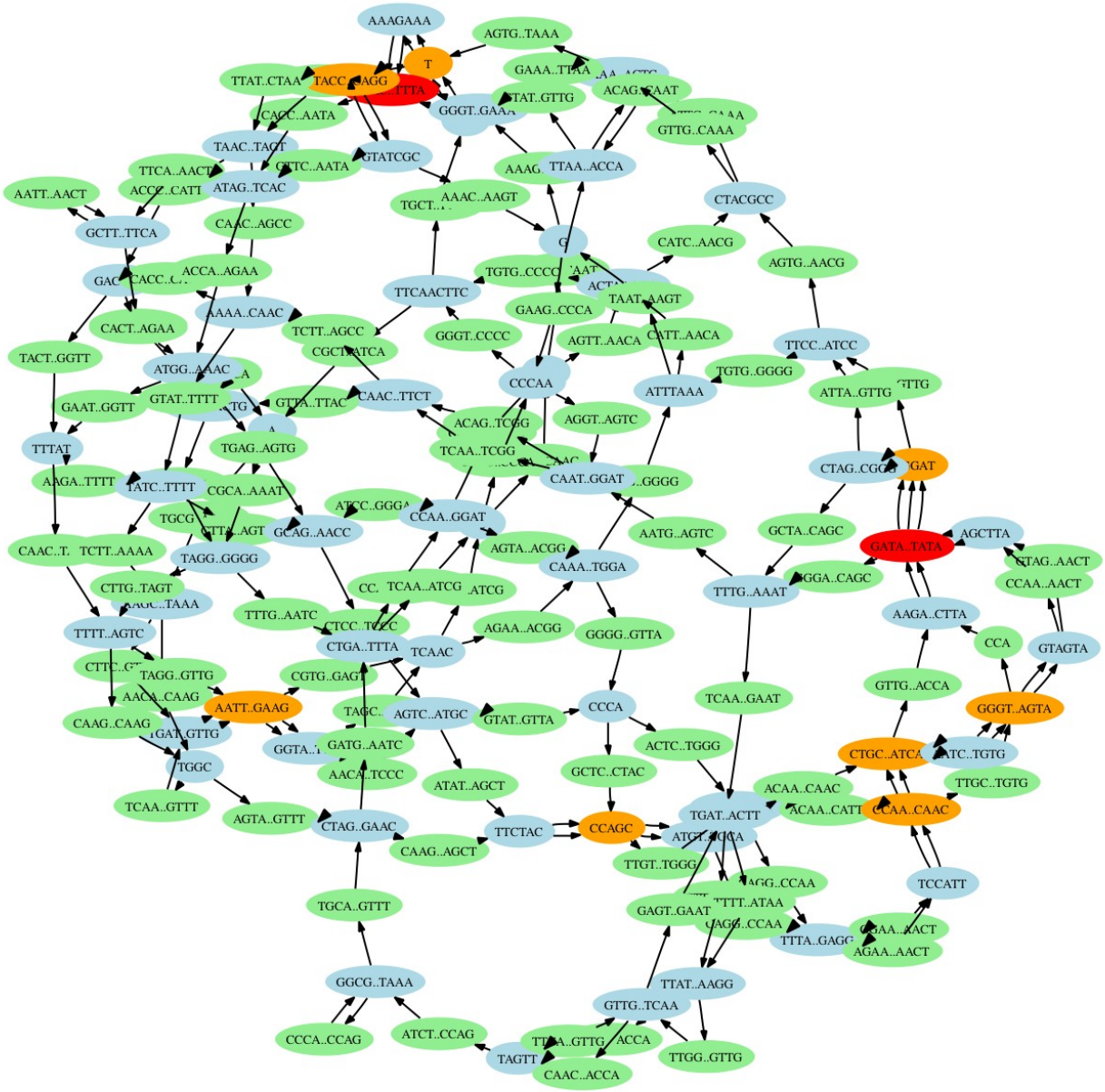
Can we do better than greedy?



Graph-based assembly

- Better than Greedy (can see better what is going on)

- Repeats still a problem


AAAGAAA
AGTG..TAAA
T
TTAT..CTAA
TACC..AAGG
GAAA..TTAA
ACAG..GAAT
GGGT..GAAA
GTTG..CAAA
TAAC..TAGT
GTATCGC
TTCA..AACT
GTTC..AATA
AAAG..TTAA..ACCA
AATT..AACT
ATAG..TCAC
AAAC..AAGT
CTACGCC
CAAC..AGCC
GCTT..TTCA
CATC..AACG
AGTG..AACG
ACCA..AGAA
TGTG..CCCC
TTCAACTTC
GAAG..CCCA
TAAT..AAGT
CACT..AGAA
AAAA..CAAC
TCTT..AGCC
GGGT..CCCC
AGTT..AACA
CATT..AACA
TACT..GGTT
CGCT..ATCA
TTCC..ATCC
ATGG..AAAC
CCCAA
TGTG..GGGG
ATTTAAA
ATTA..GTTG
GAAT..GGTT
GTAT..TTTT
GTTA..TTAC
CAAC..TTCT
ACAG..TCGG
AGGT..AGTC
TTTAT
TGAG..AGTG
TCAA..TCGG
AAGA..TTTT
TATC..TTTT
CGCA..AAAT
CAAT..GGAT
CTAG..CGGA
GAT
ATCC..GGGA
CCAA..GGAT
GCTA..CAGC
AATG..AGTC
TGCG
GTTA..AGT
GCAG..AACC
GATA..TATA
AGCTTA
CAAC..T..
TCTT..AAAA
AGTA..ACGG
GTAG..AACT
TAGG..GGGG
CAAA..TGGA
GGGA..CAGC
CCAA..AACT
CTTG..TAGT
CC..TCAA..ATCG
TTTG..AAAT
AAGA..CTTA
AAGC..TAAA
TTTG..AATC
CCA
TTTT..AGTC
AGAA..ACGG
GGGG..GTTA
GTAGTA
CTGA..TTTA
TCAAC
CTTC..GT..
TAGG..GTTG
CGTG..GAGT
TCAA..GAAT
GTTG..ACCA
AACA..CAAG
AATT..GAAG
TAGC..
AGTC..ATGC
GTAT..GTTA
CCCA
GGGT..AGTA
CAAG..CAAG
GAT..GTTG
GGTA..
GATG..AATC
ACTC..TGGG
CTGC..ATCA
ATC..TGTG
TGGC
AACA..TCCC
GCTC..CTAC
TTGC..TGTG
TCAA..GTTT
ATAT..AGCT
ACAA..CAAC
AGTA..GTTT
CTAG..GAAC
TTCTAC
CCAGC
TGAT..ACTT
ACAA..CATT
CCAA..CAAC
CAAG..AGCT
TTGT..TGGG
TCCATT
TGCA..GTTT
GAGT..GAAT
TTTT..ATAA
AAGG..CCAA
TTTA..GAGG
AGAA..AACT
GGCG..TAAA
TTAT..AAGG
GTTG..TCAA
CCCA..CCAG
ATCT..CCAG
TTAA..GTTG
ACCA
TAGTT
TTGG..GTTG
CAAC..ACCA

Next: Assembly paradigms