Read Mapping

de Bruijn Graph-based de novo Assembly

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

Genomics: Lecture #3 WS 2014/2015

Today

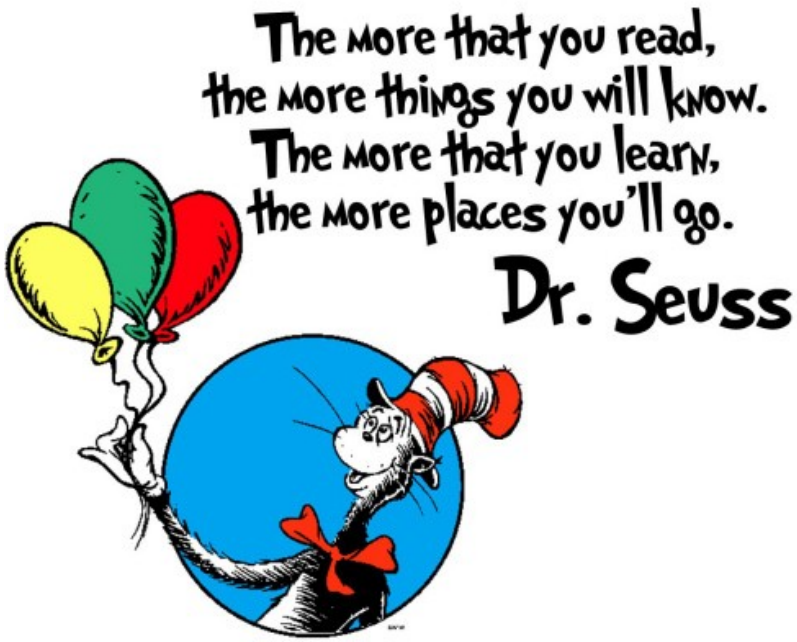
Last time we looked at some basic concepts of genome sequencing and assembly. It should however be clear that just looking for a Eulerian path in a de Bruijn graph will not solve all problems. Today, we will look at some ideas and concepts to use de Bruijn graphs for practical assembly algorithms

Main source for today:

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *PNAS* **98**:9748-53.

The Dr. Seuss-Ome

Just to review the topics of the last lecture, we will adapt a brilliant idea of Michael Schatz (CHSL), who used the first sentence of Dicken's *A Tale of Two Cities* to illustrate De Bruijn graphs



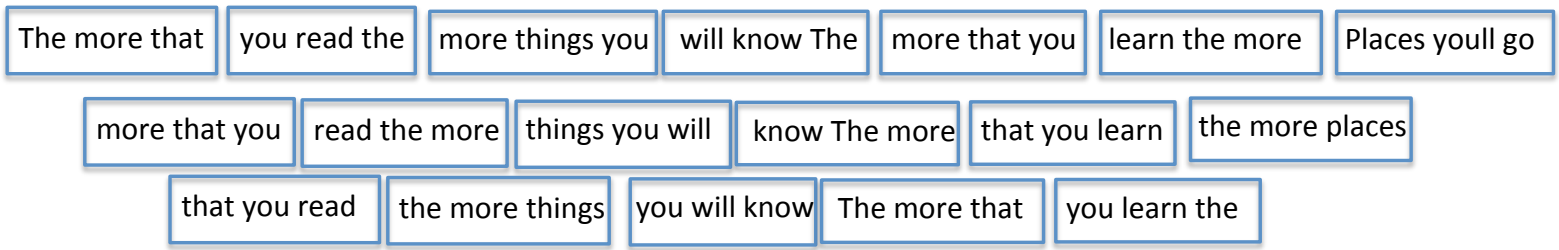
Dr. Seuss (Theodor Seuss Geisel), American writer of children's books. *I Can Read With My Eyes Shut!* (1978)

The Dr. Seuss-Ome

Shredded Seuss Reconsruction

Imagine Seuss accidentally shreds the first printing of *I can read with my eyes shut!*

The more that you read, the more things you will know. The more that you learn, the more places you'll go.



- Many different copies of the book are shredded into three word fragments ("3-mer" subsequences)
- Start position of the fragments is random
- Goal: find overlaps to reconstruct the Seussome

The Dr. Seuss-Ome

know The more
learn the more
more that you
more that you
more things you
Places youll go
read the more
that you learn
that you read
The more that
The more that
the more places
the more things
things you will
you learn the
you read the
you will know
will know The

Greedy Reconsruction

- Let‘s try to reconstruct the original text on the basis of overlaps
- Start with an arbitrary fragment
- „Extend“ the fragment with fragments whose 2-prefix matches the last 2-suffix

the more things

We choose this 3-mer „at random“

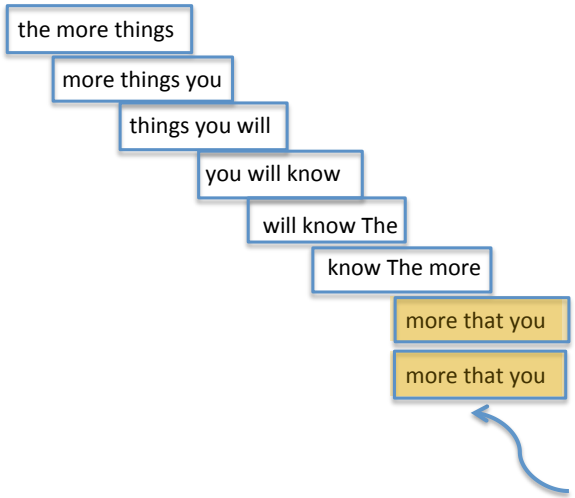
The Dr. Seuss-Ome

Peter N. Robinson

know The more
learn the more
more that you
more that you
more things you
Places youll go
read the more
that you learn
that you read
The more that
The more that
the more places
the more things
things you will
you learn the
you read the
you will know
will know The

Greedy Reconsruction

- „Extend“ the fragment with fragments whose 2-prefix matches the last 2-suffix



Which 3-mer should we extend now??

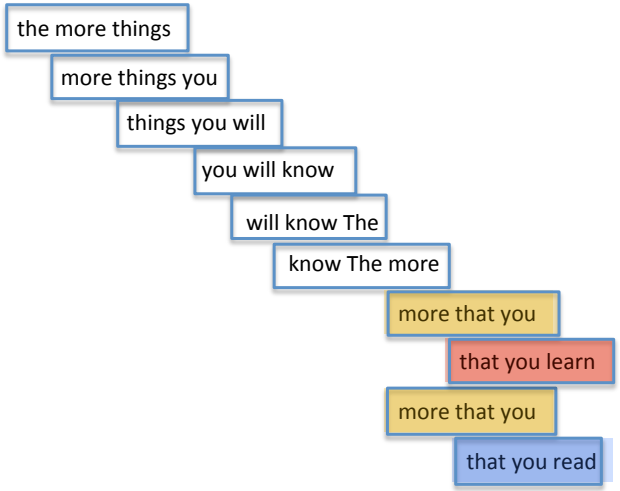
The Dr. Seuss-Ome

Peter N. Robinson

know The more
learn the more
more that you
more that you
more things you
Places youll go
read the more
that you learn
that you read
The more that
The more that
the more places
the more things
things you will
you learn the
you read the
you will know
will know The

Greedy Reconsruction

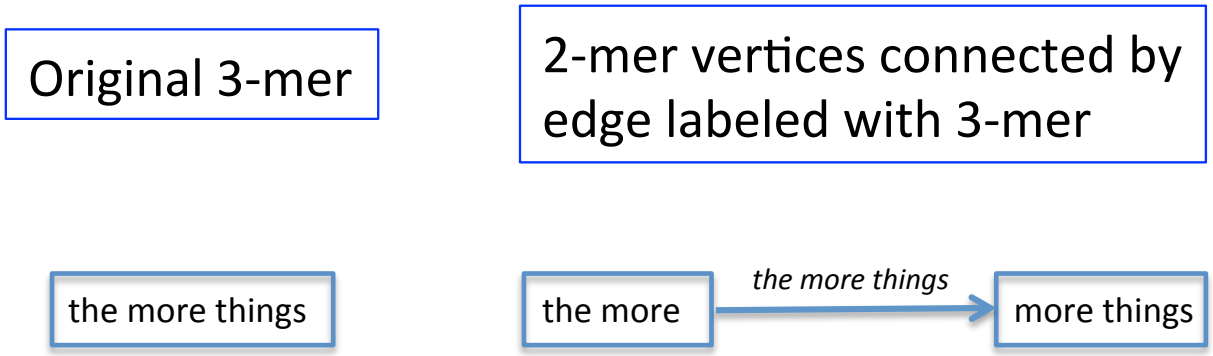
- „Extend“ the fragment with fragments whose 2-prefix matches the last 2-suffix



The repeated k-mer makes the correct reconstruction ambiguous

The more things you will know the more that you [***learn/read***]

The Dr. Seuss-Ome



- $G = (V, E)$
- V all length $k-1$ fragments (here: $k-1=2$)
- E directed edges between consecutive subfragments with labels of length k (here: $k=3$)
- Note that vertices overlap by $k-2$ words (here: $k-2=1$)

The Dr. Seuss-Ome

- How do we choose a value for k in real life?
- "Big enough": the $k-1$ mer sequences should mainly be unique
- However, memory usage grows as $\mathcal{O}(nk)$, or about $n \approx 2.4 \times 10^9$ nucleotides, k-mer size $k = 27$, requiring about 15 GB (nk/4 bytes) of memory to store the nodes alone.
- Repeats in typical genomes are larger than individual reads, so even if we could hope to sequence without errors, we would not quite yet have a complete solution to the assembly problem

We will now construct a de Bruijn graph (DBG) from the Seussome as explain in the previous lecture

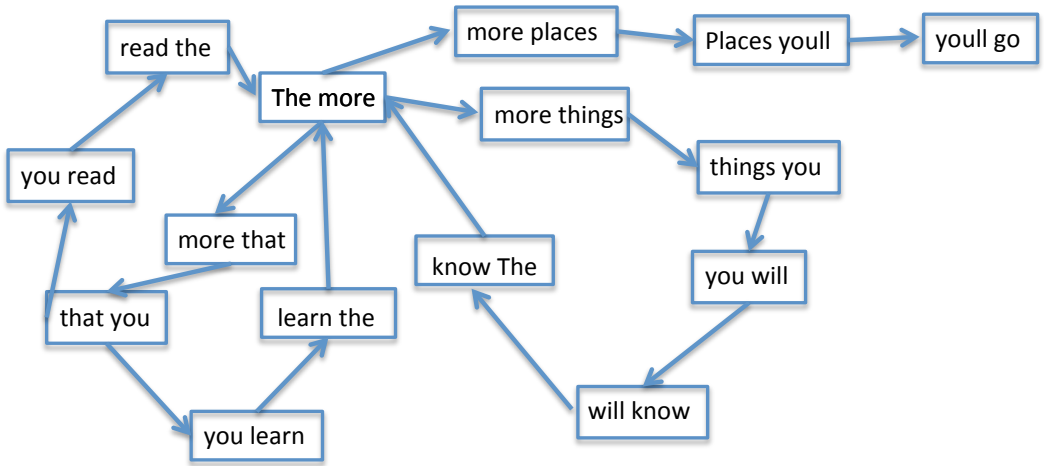
The Dr. Seuss-Ome

Peter N. Robinson

know The more
learn the more
more that you
more that you
more things you
Places youll go
read the more
that you learn
that you read
The more that
The more that
the more places
the more things
things you will
you learn the
you read the
you will know
will know The

DBG Reconstruction

The more that you read, the more things you will know. The more that you learn, the more places you'll go.



A particular Eulerian tour of the graph reconstructs the original text but there are multiple such tours

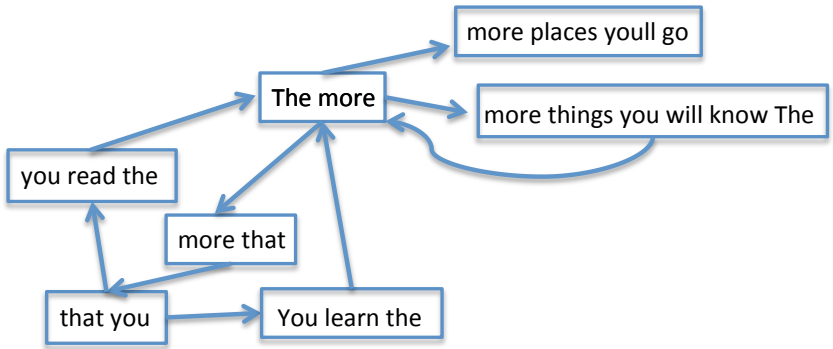
The Dr. Seuss-Ome

Peter N. Robinson

know The more
learn the more
more that you
more that you
more things you
Places youll go
read the more
that you learn
that you read
The more that
The more that
the more places
the more things
things you will
you learn the
you read the
you will know
will know The

DBG Compression

The more that you read, the more things you will know. The more that you learn, the more places you'll go.



After reconstruction, many edges are unambiguous and can be compressed

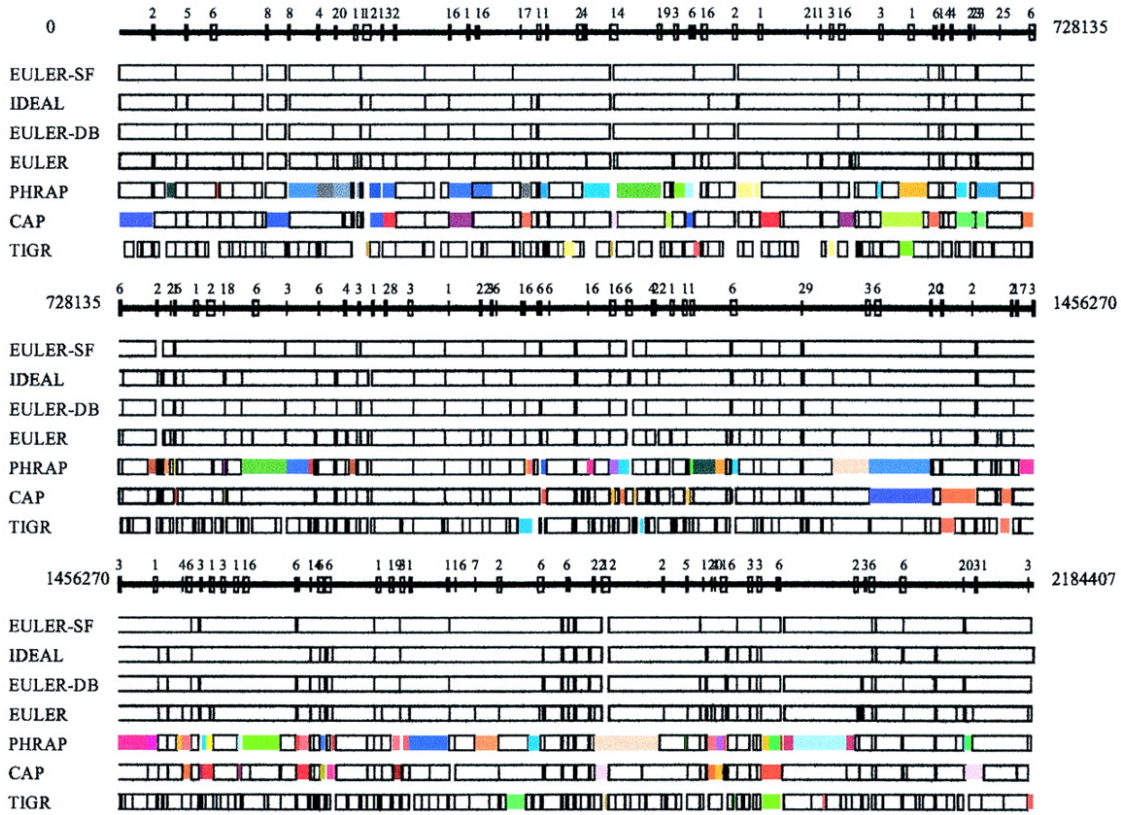
Back to real life...

Unfortunately, there is a time when it is necessary to wake up to the realities of real life...In the rest of the lecture we will cover the EULER algorithm as developed by Pevzner, Tang and Waterman (2001).

The authors showed that EULER was superior to previous algorithms based the overlap-layout-consensus paradigm. There were two key ideas

- Intelligent way of dealing with repetitive regions
- Intelligent way of dealing with sequence errors
- Superpath approach to disambiguating the de Bruijn graph

EULER



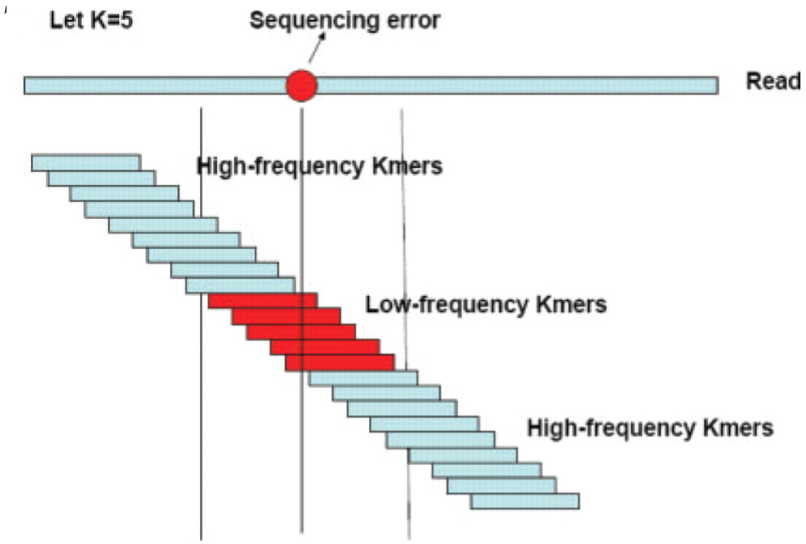
Comparative analysis of euler, phrap, cap, and tigr assemblers

- Every box corresponds to a contig in *Neisseria meningitidis* assembly
- colored boxes correspond to assembly errors

EULER

Unfortunately, the straightforward Eulerian path approach, although very promising, did not scale up well. The problem is that sequencing errors transform a simple de Bruijn graph into a tangle of erroneous edges.

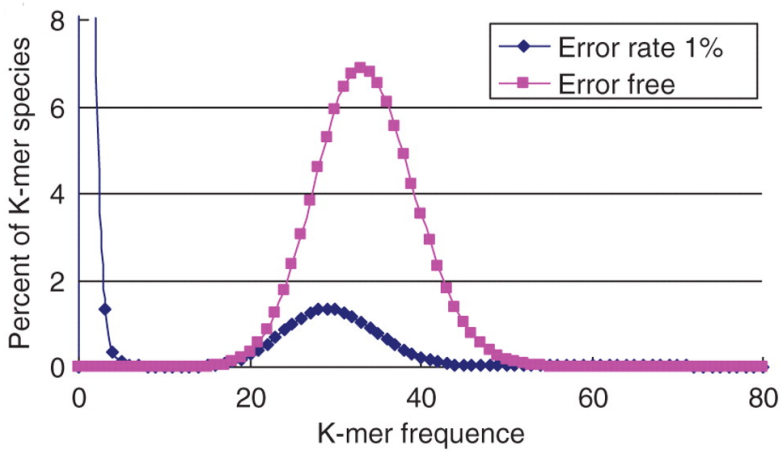
EULER



- Li et al. (2011) Briefings in functional genomics (2012) 11 (1): 25-37.

- The 5 k-mers which crossed the error base appear in low frequency
- the surrounding k-mers appear in high frequency.
- In practice, the situations are often more complex than this

EULER



- Li et al. (2011) Briefings in functional genomics (2012) 11 (1): 25-37.

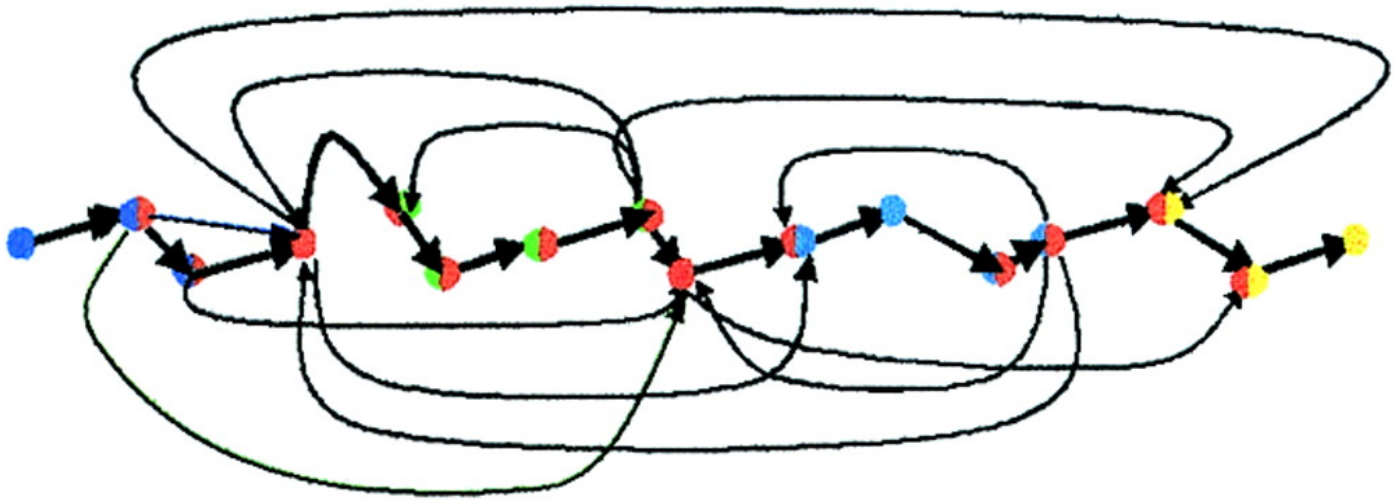
- Distribution of 17-mer frequency for error free and 1% erroneous data
- In the 1% error curve, about 80% k-mer species have frequency below five, most of which are caused by sequencing errors.

Therefore, an obvious heuristic is to remove low-frequency k-mers from the assembly

EULER

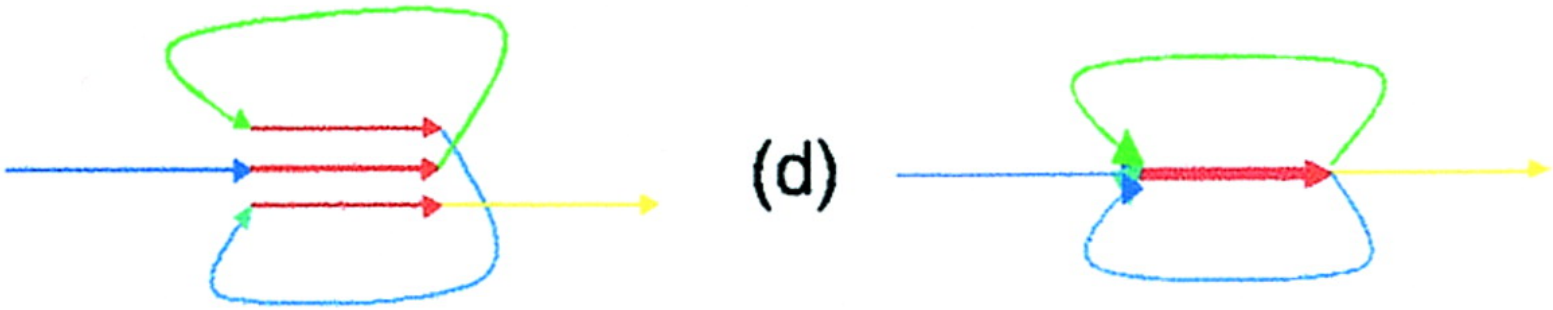
- Consider this DNA sequence, which consists of four unique segments A, B, C, D, and one triple repeat R (in red).
- We perform WGS and obtain 16 reads
- The reads are here conveniently colored according to their sequence of origin

EULER



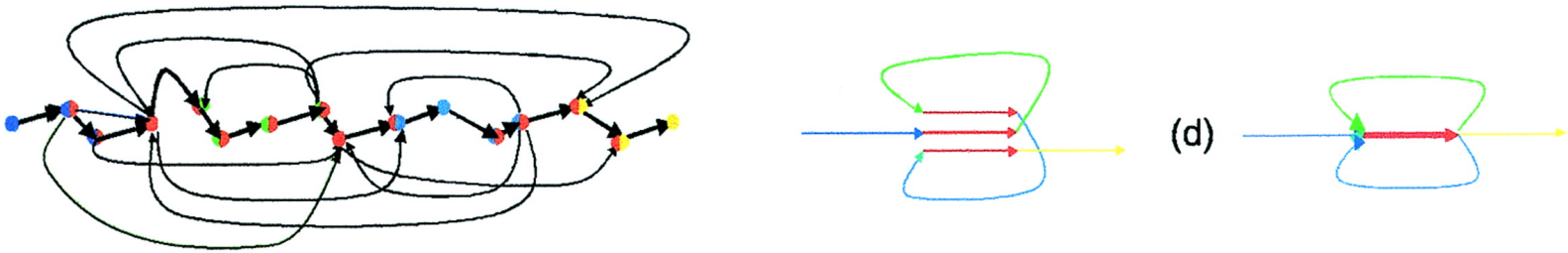
- Every read corresponds to a vertex in the **overlap graph**
- two vertices are connected by an edge if the corresponding reads overlap.
- The fragment assembly problem is thus cast as finding a path in the overlap graph visiting every vertex exactly once, a **Hamiltonian Path Problem**.
- The Hamiltonian Path Problem is NP-complete
- This is why fragment assembly of highly repetitive genomes is a notoriously difficult problem.

EULER



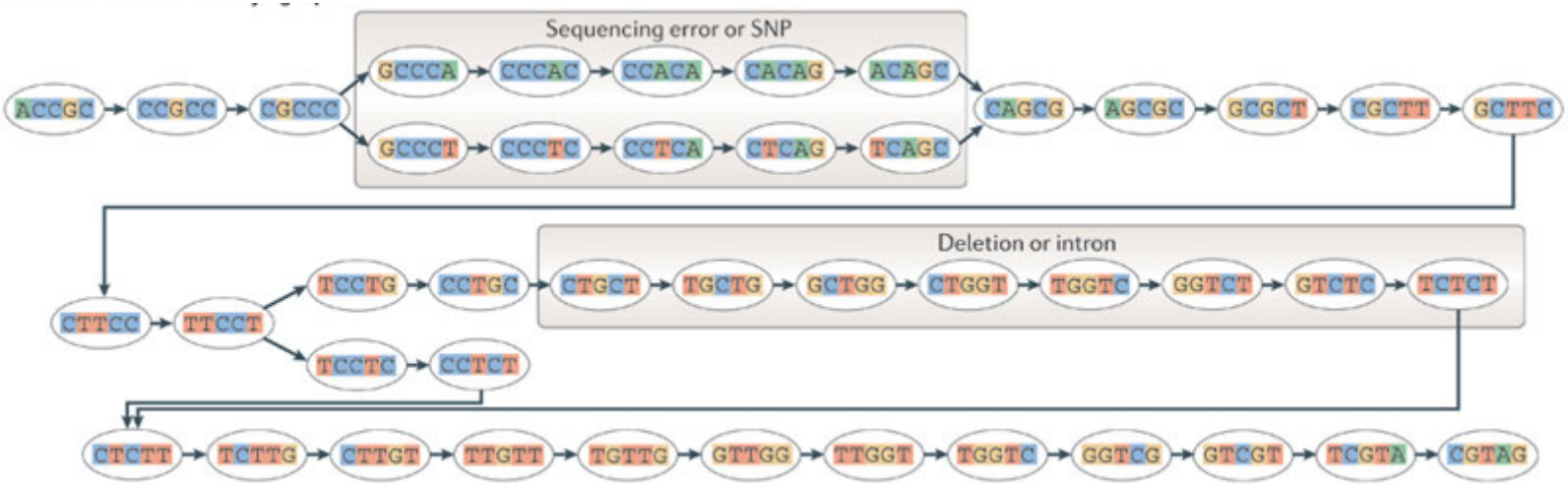
- In an informal way, one can visualize the construction of the de Bruijn graph by representing a DNA sequence as a "thread" with repeated regions covered by a "glue" that "sticks" them together
- The resulting de Bruijn graph consists of $4 + 1 = 5$ edges (we assume that the repeat edge is obtained by gluing three repeats and has multiplicity three).
- In this approach, every repeat corresponds to an edge rather than a collection of vertices in the layout graph.

EULER



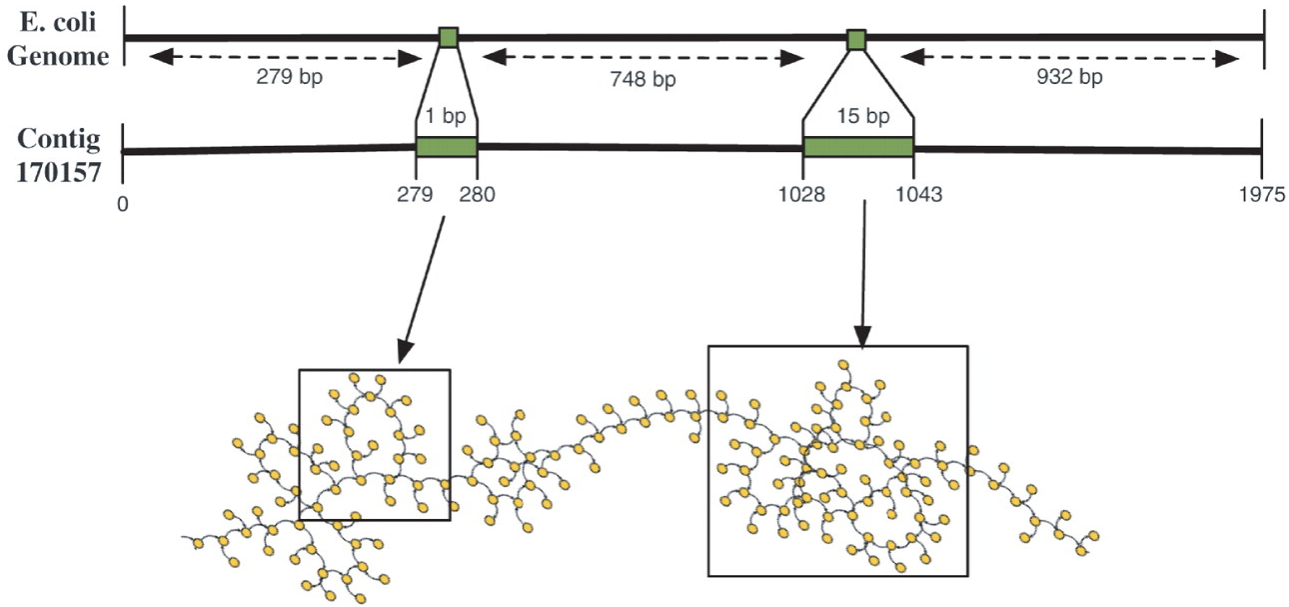
- Obviously, de Bruijn graph is a much simpler representation of repeats than the overlap graph
- fragment assembly is now cast as finding a path visiting every edge of the graph exactly once, an Eulerian Path Problem.
- There are two Eulerian paths in the graph: one of them corresponds to the sequence reconstruction ARBRCRD, whereas the other one corresponds to the sequence reconstruction ARCRBRD.
- In contrast to the Hamiltonian Path Problem, the Eulerian path problem is easy to solve even for graphs with millions of vertices, because there exist linear-time Eulerian path algorithms

EULER



- Read errors cause characteristic patterns in the de Bruijn graph
- The numer of nodes “explodes”

EULER



Ronen R (2012) Bioinformatics 28:i188-96.

- Up to 90% of nodes in de Bruijn assembly grphs may stem from sequence errors
- Here: The alignment of a 1975 bp contig from the assembly with Velvet and k=31, showing two insertions in the alignment, having respective lengths 1 bp and 15 bp.
- The de Bruijn graph constructed from the set of permissively aligned reads to this contig contains bulges and whirls at regions corresponding to the insertions in the contigs

EULER: error correction

- If we knew the genome sequence and could somehow correctly align the reads, it would be relatively easy to perform error correction
- But of course we do not know the true genome seqeunce G
- If we knew the set of all k-mers present in G, we could also try to correct the reads accordingly
- We can approximate this set based on the evidence in the reads data using the assumption that true k-mers are present multiple times in the collection of reads but error-related k mers have low counts in the data

$\boxed{G_k}$: the set of all k-tuples in genome G

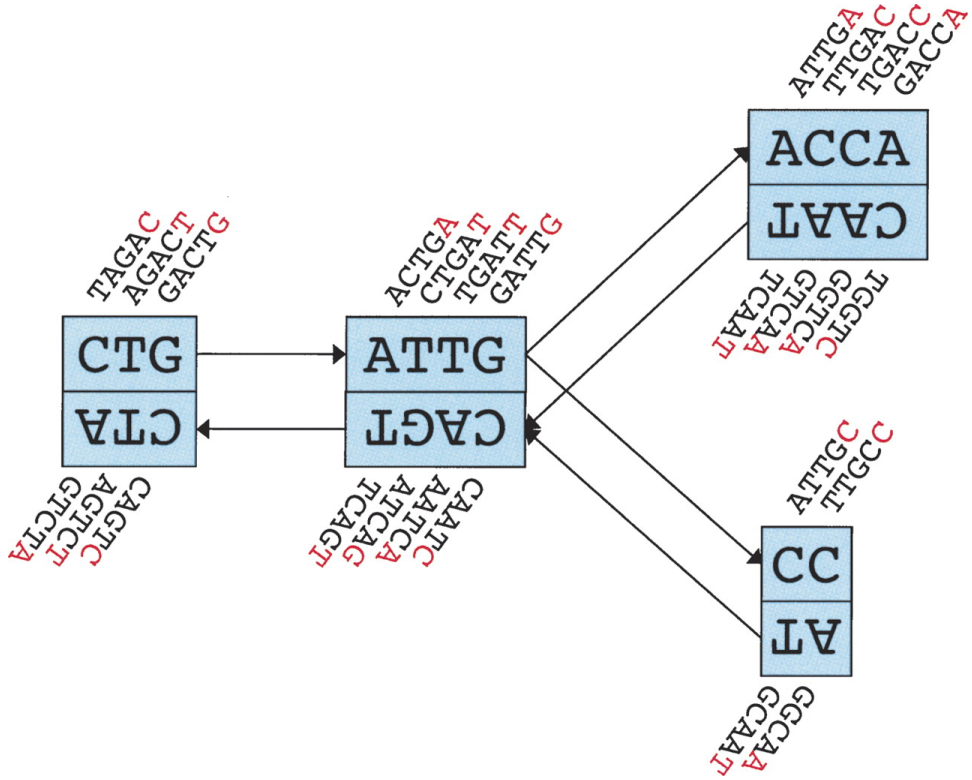
EULER: error correction

$\boxed{G_k}$: the set of all k-tuples in genome G

- An k-mer is called **solid** if it belongs to more than M reads and **weak** otherwise.
- M is a threshold, e.g., 4
- A natural approximation for G_k is the set of all solid k-mers from a sequencing project.
- Now let T be a collection of k-mers called a **spectrum**.
- A string s is called a T-string if all its k-mers belong to T.

EULER: Strand ambiguity

- WGS is typically not strand-specific
- Solution: Treat each k-mer as actually being two k-mers, the original sequence and the reverse complement



EULER: error correction

Spectral Alignment Problem.
Given a string s and a spectrum T, find the minimum number of mutations in s that transform s into a T-string.

We are given a collection of reads (strings) $S = \{s_1, \ldots, s_n\}$ from a sequencing project and an integer k.

- The **spectrum** of S is a set S_k of all k-mers from the reads $s_1, \ldots, s_n$ and $\overline{s}_1, \ldots, \overline{s}_n$
- $\overline{s}$ denotes a reverse complement of read s.
- Let Δ be an upper bound on the number of errors in each DNA read

EULER: error correction

Error Correction Problem.
Given S, Δ, and k, introduce up to Δ corrections in each read in S in such a way that $|S_k|$ is minimized.

- An error in a read s affects at most k individual k-mers in s and k more k-mers in $\overline{s}$
- Thus, an error usually creates $2k$ erroneous k-mers that point to the same sequencing error
- If the error is close to the end of a read, less erroneous k-mers are created ($2d$ for positions within a distance $d < k$ from the endpoint of the reads)
- A greedy approach for the Error Correction Problem is to look for error corrections in the reads that reduce the size of S_k by $2k$ (or $2d$ for positions close to the endpoints). This simple procedure already eliminates 86.5% of the errors in sequencing reads.

EULER: error correction

Greedy Error Correction.
Look for error corrections in the reads that reduce the size of S_k by $2k$ (or $2d$ for positions close to the endpoints).

- Why does this work?
- This approach eliminated 86.5% of errors in the test bacterial genomes used
- Similar heuristics improve error correction to 98% (look for k-mers that are nearly identical to other k-mers with high multiplicity)
- For details see Pevzner PA, Tang H, Waterman MS (2001) A new approach to fragment assembly in DNA sequencing. In Proceedings of the Fifth International Conference on Computational Biology (RECOMB 2001, Montreal). pp. 256265.