(VERSIONE 1 - ult. agg. 6.12.2025)

# PROPOSTE DI PROGETTO PER L'ESAME (R. ZIZZA)

INSEGNAMENTO: STRUMENTI FORMALI PER LA BIOINFORMATICA, A.A. 2025-26

## INTRODUZIONE

Questo documento contiene l'elenco dei progetti proposti per lo svolgimento della prova d'esame, se si desidera svolgerla sugli argomenti della seconda (e parzialmente prima) parte del corso.

La lista non è esaustiva: è una proposta. Altri argomenti sono stati proposti durante le lezioni, sia nella prima sia nella seconda parte del corso (vedere slides), anche in relazione ad argomenti puramente teorici. Inoltre, *gli studenti stessi possono proporre argomenti che intendono conoscere e/o approfondire.*

Ogni progetto prevede:

1. comprensione del problema generale
2. lettura della bibliografia indicata per lo specifico progetto assegnato
3. studio approfondito dell'articolo selezionato
4. eventuale studio del codice
5. eventuale esecuzione del tool selezionato e testing su dati genomici (concordati con il docente, se non reperibili attraverso l'articolo)

Viene sollecitato l'uso della piattaforma **Galaxy** (https://usegalaxy.eu/) per l'esecuzione dei tool, se possibile, o dei vari Genome Browser.

Dopo aver selezionato un progetto, le specifiche di sviluppo saranno concordate con la Prof.ssa Zizza. Il progetto poi dovrà essere presentato alla classe (presentazione Powerpoint/Beamer) e accompagnato da una breve relazione/documentazione scritta di supporto, che spieghi il progetto selezionato e il lavoro svolto. Se si tratta di lavoro originale di ricerca, le specifiche saranno concordate con il docente.

**Composizione gruppo**: eventualmente è possibile svolgere il progetto in gruppo (massimo 3). Ovviamente la difficoltà del lavoro cresce in base al numero di componenti. Chi sceglie di svolgere il progetto d'esame su questa seconda parte deve inviate e-mail a rzizza@unisa.it, scrivendo

- ✓ OGGETTO: progetto esame SFB
- ✓ corpo del messaggio: nome/cognome dei partecipanti, scelta del progetto

Si riceverà una e-mail con la data di colloquio per fissare le specifiche del lavoro da svolgere.

**Data d'esame:** si auspica che l'esame si concluda nella sessione invernale. Gli incontri di presentazione saranno schedulati in gruppi e comunicati, dopo aver raccolto le vostre disponibilità.

**Voto:** come spiegato, se si aspira ad un voto superiore al 28, occorre effettuare uno studio molto approfondito sulla struttura algoritmica del lavoro.

**Nota bene:** gli argomenti indicati con "*" sono quelli in cui è richiesta attività di ricerca, intesa come sviluppo di nuove tecniche teoriche/pratiche

# ALLINEAMENTO

## PROGETTO 1: ALLINEAMENTO A COPPIE / MULTIPLO

Viene riportato un elenco di alcune tecniche e tool associati.

- Space efficiente sequence alignment (dal libro di Pevzner)

- Studio di BLAST https://pubmed.ncbi.nlm.nih.gov/2231712/

- PSI-BLAST https://academic.oup.com/nar/article/25/17/3389/1061651

- [survey] An overview of Multiple Sequence Alignnments and Cloud Computing in Bioinformatics https://onlinelibrary.wiley.com/doi/10.1155/2013/615630

- T-Coffee https://pmc.ncbi.nlm.nih.gov/articles/PMC3125728/

- Muscle https://pmc.ncbi.nlm.nih.gov/articles/PMC390337/

- [importanza dell'albero guida per MSA finale]
  Simple chained guide trees give high-quality protein multiple sequence alignments https://www.pnas.org/doi/10.1073/pnas.1405628111

- [importanza dell'albero guida per MSA finale]
  The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses https://pubmed.ncbi.nlm.nih.gov/18229674/

## PROGETTO 2: PROFILE REPRESENTATION OF MULTIPLE ALIGNMENT

Un elenco di vari approfondimenti, degli articoli e/o dei tool, con la produzione di benchmark di confronto (sul sito dell'EMBL-EBI), usando HMM.

https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/ Multiple+Sequence+Alignment)

- Studio dell'utilizzo delle Hidden Markov Models & bioinformatica
  - Libro di Durbin, nella cartella relativa sul sito del corso.
  - HMM and their applications in biological sequence analysis https://pmc.ncbi.nlm.nih.gov/articles/PMC2766791/

- HMMER: Eddy, S. R. 2001. HMMER: profile hidden Markov models for biological sequence analysis. http://hmmer.wustl.edu

- *Relazione tra Hidden Markov Models e automi probabilistici.
  [argomento da sviluppare completamente! Si tratta di un'idea di ricerca]

- HMM e la regola 11/25: https://pubmed.ncbi.nlm.nih.gov/11907225/

- Uso di progressive alignment in ClustalW (https://www.sciencedirect.com/science/article/abs/pii/0378111988903307?via%3Dihub) e ruolo di HMM in Clustal Omega (https://link.springer.com/protocol/10.1007/978-1-62703-646-7_6).

- Clustal Omega https://pmc.ncbi.nlm.nih.gov/articles/PMC5734385/

- mBed per Clustal Omega: Sequence embedding for fast construction of guided trees for multiple sequence alignments https://almob.biomedcentral.com/articles/10.1186/1748-7188-5-21

- *Relazione tra allineamento multiplo e espressioni regolari - stringhe degeneri
  [argomento da sviluppare completamente! Si tratta di un'idea di ricerca]

## PROGETTO 3: TECNICHE ALIGNMENT-FREE

Analisi e sperimentazione di tool di confronto tra sequenze senza allineamento

- [survey] https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1319-7

- [Uso del minimizer per allineare] MinHash Alignment Process (MHAP)

- Local sensitive hashing (LSH) for the edit distance

- [progetto connesso alla terza parte del corso] Locality-Sensitive Hashing-Based k-Mer Clustering for Identification of Differential Microbial Markers Related to Host Phenotype. https://pmc.ncbi.nlm.nih.gov/articles/PMC9464365/

# SEQUENZIAMENTO E ASSEMBLAGGIO

## PROGETTO 4: ASSEMBLY ALGORTHMS - OVERLAP DETECTION

Suvey di confronto tra i due approcci:

- Comparison of the two major classes of assembly algorithms: https://pubmed.ncbi.nlm.nih.gov/22184334/
- Graph Theoretical Strategies in Denovo Assembly https://ieeexplore.ieee.org/abstract/document/9684373
- Current challenges and solutions of de novo assembly https://link.springer.com/article/10.1007/s40484-019-0166-9

- Primi lavori su OLC, fino a "The independence of our genome assembly" https://pmc.ncbi.nlm.nih.gov/articles/PMC152237/

- Edena https://genome.cshlp.org/content/18/5/802.full.pdf+html

- String Graph (uso di FM-index e BWT) https://pmc.ncbi.nlm.nih.gov/articles/PMC3290790/

- Canu https://genome.cshlp.org/content/27/5/722

- Linear time complexity de novo long read genome assembly with GoldRush https://www.nature.com/articles/s41467-023-38716-x
- [Scaffolding] Overlap Graph for Assembling and Scaffolding Algorithms
- Bechmarking long read assembly https://www.sciencedirect.com/science/article/pii/S2215017X2500058X
- [Tool di integrazione dei due approcci] Integration of String and de Bruijn graphs for genome assembly https://academic.oup.com/bioinformatics/article/32/9/1301/1744507

- QUAST: quality assessment tool for genome assemblies
  https://pubmed.ncbi.nlm.nih.gov/23422339/

- Overlap detection : FODI
  https://www.sciencedirect.com/science/article/pii/S1476927125002373
- * LROD: An Overlap Detection Algorithm for Long Reads Based on k-mer Distribution
  https://www.frontiersin.org/articles/10.3389/fgene.2020.00632/full

  *Si tratta di continuare un progetto già iniziato atto a migliorare il tool pubblicato*

## PROGETTO 5: RAPPRESENTAZIONE E USO DEI DE BRUIJN GRAPH PER L'ASSEMBLAGGIO

A partire dalla survey suggerita, studiare e confrontare tool per la rappresentazione succinta dei grafi di de Bruijn, come BOSS indicato qui.

Abstract: *High-throughput sequencing has become an increasingly central component of microbiome research. The development of de Bruijn graph-based methods for assembling high-throughput sequencing data has been an important part of the broader adoption of sequencing as part of biological studies. Recent advances in the construction and representation of de Bruijn graphs have led to new approaches that utilize the de Bruijn graph data structure to aid in different biological analyses…*

- Euler: https://pmc.ncbi.nlm.nih.gov/articles/PMC55524/
- Velvet: https://pmc.ncbi.nlm.nih.gov/articles/PMC2952100/
- AllPaths: https://pmc.ncbi.nlm.nih.gov/articles/PMC2336810/
- Abyss: https://pmc.ncbi.nlm.nih.gov/articles/PMC2694472/
  (e sue versioni successive)

- Minia: Space-efficient and exact de Bruijn graph representation based on a Bloom filter
  https://almob.biomedcentral.com/articles/10.1186/1748-7188-8-22
- SPAdes: a new genome assembly algorithm annd its application to Single-Cell sequencing https://pubmed.ncbi.nlm.nih.gov/22506599/

- Bowe, A., Onodera, T., Sadakane, K., Shibuya, T. (2012). Succinct de Bruijn Graphs. In: Raphael, B., Tang, J. (eds) Algorithms in Bioinformatics. WABI 2012. Lecture Notes in Computer Science, vol 7534, pp. pp 225–235. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33122-0_18

- Edge minimization in de Bruijn graphs
- Eliminazione delle bolle nei dBG

- Space efficient merging of succint de Bruijn graphs (uso della BWT per gestire i kmer e quindi costruire dBG)

- [survey] Applications of de Bruijn graphs in microbiome research, Keith Dufault Thompson, Xiaofang Jiang, First published: 01 March 2022 https://doi.org/10.1002/imt2.4

- [progetto connnesso alla terza parte del corso] Graph Neural Network Meets de Bruijn Genome Assembly

## PROGETTO 6: ASSEMBLATORI BASATI SU KMER

Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937194/

# STRUTTURE DATI

## PROGETTO 7: BLOOM FILTERS - KMER COUNTING

Scopo: Studio dei Bloom Filter. Confronti dei vari tool che implementano Bloom Filter e che li usando per indicizzare kmer per sequenze genomiche. Replica dei test effettuati nei lavori.

Abstract: *When indexing large collections of short-read sequencing data, a common operation that has now been implemented in several tools (Sequence Bloom Trees and variants, BIGSI) is to construct a collection of Bloom filters, one per sample. Each Bloom filter is used to represent a set of k-mers which approximates the desired set of all the non-erroneous k-mers present in the sample…"*

- Survey on Bloom Filters: https://ieeexplore.ieee.org/document/8229957
- BioBloom: https://pmc.ncbi.nlm.nih.gov/articles/PMC4816029/

- Téo Lemane, Paul Medvedev, Rayan Chikhi, Pierre Peterlongo, kmtricks: efficient and flexible construction of Bloom filters for large sequencing data collections, BIOINFORMATICS ADVANCES, Volume 2, Issue 1, 2022, vbac029, https://doi.org/10.1093/bioadv/vbac029

- S. Nayak and R. Patgiri, "A Review on Role of Bloom Filter on DNA Assembly," in IEEE Access, vol. 7, pp. 66939-66954, 2019, doi: 10.1109/ACCESS.2019.2910180.

- Using cascading Bloom filters to improve the memory usage for de Brujin graphs https://almob.biomedcentral.com/articles/10.1186/1748-7188-9-2

- Data structures to represennt k-long DNA sequences
  https://dl.acm.org/doi/10.1145/3445967

- KMC3: https://academic.oup.com/bioinformatics/article/33/17/2759/3796399

- Constructing cascade bloom filters for efficient access enforcement
  https://www.sciencedirect.com/science/article/pii/S0167404818311271

- FASTK https://github.com/thegenemyers/FASTK

- A survey of k-mer methods and applications in bioinformatics
  https://pubmed.ncbi.nlm.nih.gov/38840832/
- ClassPro: Accurate k-mer Classification Using Read Profiles
  https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.WABI.2022.10
- Sparse and skew hashing of kmers
  https://academic.oup.com/bioinformatics/article/38/Supplement_1/i185/6617506

## PROGETTO 8: SUFFIX TREE, SUFFIX ARRAY, BWT

- Costruzione efficiente di queste strutture dati. Bibliografia da concordare con il docente.

- FM-index: https://dl.acm.org/doi/10.1145/1082036.1082039
- An accelerated FM-index: https://pubmed.ncbi.nlm.nih.gov/37961504/

- Costruzione di suffix array, LCP e BWT per collezione di stringhe. Lettura e comprensione dell'articolo, analisi dei tool e testing
  Louza, Felipe A. and Telles, Guilherme P. and Gog, Simon and Prezza, Nicola and Rosone, Giovanna, gsufsort: constructing suffix arrays, LCP arrays and BWTs for string collections, Algorithms Mol Biol 15, 18 (2020).

## PROGETTO 10: BWT ALIGNER E CONFRONTI

Studio di una selezione di tool di allineamento basati sulla BWT, come BWA e Bowtie (vedere le slides del corso per i riferimenti su questo progetto). Riproduzione dei test di confronto, preferibilmente usando Galaxy.

- Benchmarking short sequence mapping tools https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-184
- BWA
- Bowtie
- Bowtie2
- Costruzione efficiente della BWT

## PROGETTO 11*: FATTORIZZAZIONE DI LYNDON, SUFFIX ARRAY E BWT

1) Utilizzo della Fattorizzazione di Lyndon per la creazione efficiente del suffix array (continuazione di progetti di tesi triennale).

Abstract: *Suffix sorting is one of the most challenging question in string algorithms, aiming at building efficient data structures, too. A vast literature regards this problem and recently it has been provided an efficient technique accelerating in practice suffix sorting of a given text, by exploiting properties of Lyndon words. Our aim is to use the inverse Lyndon factorization (ICFL) of the given text, previously introduced, which factorizes the text in an increasing sequence (w.r.t. the lexicographic order) of factors, which are inverse Lyndon words. We show how we can use the suffixes of these factors (local suffixes) for inducing the sorting of the suffixes of the text. The theoretical properties on compatibility of local suffixes and bounds on the longest common prefix between two local suffixes, already proved for ICFL, can be used for suffix sorting.*

2) Definizione della BWT sulla fattorizzazione di Lyndon: completamento tool già esistente, sviluppo di test di confronto di performance con altri tool analoghi (continuazione di progetti di tesi triennale).

Abstract: *In letteratura è stata definita una variante biettiva della BWT a partire dalla Fattorizzazione di Lyndon. Si propone di proseguire con l'analisi delle perfomance di una nuova variante biettiva definita a partire da una variante della Fattorizzazione di Lyndon, recentemente introdotta. Questa gode di interessanti proprietà di limiti sulla lunghezza dell'LCP tra suffissi dei fattori della fattorizzazione introdotta.*

In entrambi i progetti si tratta di completare e ottimizzare i tool sviluppati, conducendo analisi di prestazioni complete.

In entrambe le proposte, sarebbe interessante anche uno studio teorico delle proprietà utilizzate sperimentalmente.