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

INSEGNAMENTO: STRUMENTI FORMALI PER LA BIOINFORMATICA, A.A. 2024-25

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. selezione di un tool specifico, studio del codice
4. 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. Viene anche sollecitata l'esecuzione dell'assignment caricato sul sito, relativo all'argomento scelto.

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.

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

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 MULTIPLO

Viene riportato un elenco di alcune tecniche e tool associati.

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

- [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/

- Studio delle varie tecniche di progressive alignment. Analisi e confronto di vari tool selezionati a cura dello studente.

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).

- 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

- 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

- 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

- * 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

- [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/

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…*

- 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

- 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/

  In questi progetto, oltre allo studio del tool, si può affiancare l'esecuzione di un assignment per l'assemblaggio di alcune reads (assignment di Compeau, vedi cartella sul sito).

- [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…"*

- Melsted, P., Pritchard, J.K. Efficient counting of ***k***-mers in DNA sequences using a Bloom filter. *BMC Bioinformatics* **12**, 333 (2011). https://doi.org/10.1186/1471-2105-12-333

- 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

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

- Back to sequences: A simple tool designed to index a set of *k*-mers of interests, and to stream a set of sequences, extracting those containing at least one of the indexed *k*-mer. https://www.biorxiv.org/content/10.1101/2023.10.26.564040v1.abstract

- aKmerBroom: Ancient oral DNA decontamination using Bloom filters on k-mer sets https://www.sciencedirect.com/science/article/pii/S2589004223021134X

- 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

PROGETTO 8: GSUFSORT

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 9: GALAXY

In questo progetto occorre illustrare l'utilizzo della piattaforma Galaxy su un caso di studio a scelta, come il trimming dei file FASTQ utilizzando direttamente il workflow messo a disposizione, e/o l'uso di tool di allineamento sul file ripulito.

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

ALTRI ARGOMENTI NON PRESENTATI DURANTE LE LEZIONI

PROGETTO 11: PANGENOMICA

Analisi dei grafi proposti per la pangenomica, effettuando su una selezione di essi (anche uno solo) analisi dei tool relativi e testing.

Abstract: *"The recent advances in sequencing technologies enable the assembly of individual genomes to the quality of the reference genome. How to integrate multiple genomes from the same species and make the integrated representation accessible to biologists remains an open challenge. Here, we propose a graph-based data model and associated formats to represent multiple genomes while preserving the coordinate of the linear reference genome…"*

- Alcuni tool https://pangenome.github.io/

- Canale YouTube scuola di dottorato PANGAIA project https://www.youtube.com/@pangaiaproject9687/featured

- Survey: Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E. *Pangenome Graphs*. Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139-162. doi: 10.1146/annurev-genom-120219-080406. Epub 2020 May 26. PMID: 32453966; PMCID: PMC8006571.

- Survey: Baaijens, J.A., Bonizzoni, P., Boucher, C. *et al.* Computational graph pangenomics: a tutorial on data structures and their applications. *Nat Comput* **21**, 81–108 (2022). https://doi.org/10.1007/s11047-022-09882-6

- Andrea Guarracino, Simon Heumos, Sven Nahnsen, Pjotr Prins, Erik Garrison, **ODGI**: understanding pangenome graphs, *Bioinformatics*, Volume 38, Issue 13, 1 July 2022, Pages 3319–3326, https://doi.org/10.1093/bioinformatics/btac308

PROGETTO 12: ALBERI EVOLUTIVI

Durante le lezioni non è stato possibile affrontare questa tematica, se non nell'ambito del MSA. Potrebbe essere interessante proporre questo argomento, partendo dall'assignment pratico di Compeau inserito nella cartella Assignment (da eseguire con l'uso della piattaforma MEGA) e poi completare con l'approfondimento teorico.

Chi decidesse di scegliere questa tematica, stabilirà con il docente l'assignment da eseguire.

PROGETTO 13*: 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.

PROGETTO 14*: FATTORIZZAZIONE DI LYNDON E MINHASH

Nell'ambito delle tecniche alignment-free per il confronto di sequenze genomiche, questo progetto analizza l'utilizzo di opportune rappresentazioni delle fattorizzazioni di Lyndon come misura della similarità.

Abstract: *I risultati teorici hanno mostrato come le k-fingers (k-mer degli sketch della fattorizzazione di Lyndon di una stringa), permettano di individuare le informazioni mantenute tra le fingerprints. Questa proprietà è determinante nel calcolo dell'overlap tra reads. In un lavoro di tesi triennale viene proposto uno studio sull'utilizzo della Fattorizzazione di Lyndon e sue varianti sulla capacità delle k-fingers, in combinazione con l'indice di Jaccard, di catturare similarità tra sequenze genomiche.*

Si propone di proseguire il lavoro, considerando varie ottimizzazioni già indicate (calcolo dell'indice di similarità di Jaccard eseguito su insiemi già preventivamente ordinati, sviluppo di un algoritmo ad hoc basato sul MinHash per queste tecniche combinate, analisi più esaustiva, anche tramite tecniche di Machine Learning, delle configurazioni ottimali che massimizzino la qualità informativa mantenuta nelle fingerprint).

Interessante sarebbe anche uno studio teorico delle proprietà utilizzate sperimentalmente.