

# Yuri Pirola

## Curriculum Vitae

E-Mail: [yuri.pirola@unimib.it](mailto:yuri.pirola@unimib.it)

Homepage: <https://algolab.eu/pirola>

### Posizioni

- Professore II Fascia (SSD: INF/01), DISCo, Univ. degli Studi di Milano-Bicocca.  
Periodo: 1 ottobre 2022–oggi.
- Ricercatore a Tempo Determinato (lett. B, SSD: INF/01), DISCo, Univ. degli Studi di Milano-Bicocca.  
Periodo: 1 ottobre 2019–30 settembre 2022.
- PTA di area Tecnica, Tecnico-Scientifica ed Elaborazione Dati (cat. D), Univ. degli Studi di Milano.  
Periodo: 4 gennaio 2016–30 settembre 2019.  
Analisi e sviluppo di sistemi informatici a supporto delle attività istituzionali dell'Ateneo.
- Assegnista di Ricerca, DISCo, Univ. degli Studi di Milano-Bicocca.  
Periodo: 1 maggio 2012–3 gennaio 2016.  
Titolo: “Metodi algoritmici per l’analisi di dati NGS (Next Generation Sequencing)” (SSD: INF/01).  
Bando competitivo con valutazione basata sul progetto di ricerca presentato dal candidato.
- Assegnista di Ricerca, DISCo, Univ. degli Studi di Milano-Bicocca.  
Periodo: 1 gennaio 2011–30 aprile 2012.  
Titolo: “Inferenza efficiente di aplotipi in popolazioni di animali da reddito mediante marcatori SNP ad alta densità” (SSD: INF/01)  
Finanziatori: Regione Lombardia e Parco Tecnologico Padano, Lodi.
- Collaborazione coordinata e continuativa su progetto di ricerca, Parco Tecnologico Padano, Lodi.  
Periodo: 22 marzo 2010–31 dicembre 2010.  
L’attività di ricerca si è inserita nel progetto “PROZOO” ed è stata coordinata da Dott. Alessandra Stella e Dott. Stefano Biffani sotto la supervisione del responsabile di progetto Dott. John Williams.

### Titoli di Studio

- Dottorato di Ricerca in Informatica (XXII Ciclo), Univ. degli Studi di Milano-Bicocca, 2010.  
Tesi: “*Combinatorial Problems in Studies of Genetic Variations: Haplotyping and Transcript Analysis*”.  
Supervisore: Prof. Paola Bonizzoni.
- Laurea Specialistica in Informatica, 110/110 e lode, Univ. degli Studi di Milano-Bicocca, 2006.  
Tesi: “*Analisi della Neutralità degli Spazi di Ricerca Booleani in Programmazione Genetica*”.  
Supervisori: Dott. Leonardo Vanneschi e Prof. Giancarlo Mauri.

## Partecipazione a Progetti di Ricerca

- Progetto PANGAIA “Pan-genome Graph Algorithms and Data Integration”  
*Periodo:* 2020–2023     *Finanziatore:* EU Horizon 2020 RISE.  
Il contributo in questo progetto riguarda lo sviluppo di algoritmi per la rappresentazione e l’analisi di insiemi di sequenze genomiche.
- Progetto CORSAI “Raman analysis of saliva from COPD patients as new biomarker”  
*Finanziatore:* ERA PerMed.
- Progetto “Modulation of anti-cancer immune response by regulatory non-coding RNAs”  
*Periodo:* 2014–2016     *Finanziatore:* Fondazione Cariplo.  
Il contributo in questo progetto riguarda lo sviluppo di un metodo per la determinazione di eventi di alternative splicing da dati di RNA-Seq e sua validazione sui dataset prodotti dagli altri partner di progetto.
- Progetto SPAC3 “Servizi smart della nuova PA per la Citizen-Centricity in Cloud”  
*Periodo:* 2014–2015     *Finanziatore:* Regione Lombardia.  
Il contributo in questo progetto riguarda lo sviluppo di un metodo per l’assistenza al processo di integrazione di schemi concettuali basato su tecniche di ILP.
- Progetto PRIN 2010/11 “Automati e linguaggi formali: aspetti matematici e applicativi”.  
*Periodo:* 2013–2016     *Finanziatore:* MIUR.  
Il contributo in questo progetto riguarda lo sviluppo di algoritmi basati su tecniche di indicizzazione succinta per l’analisi e il confronto efficiente di sequenze nucleotidiche ottenute con tecnologie di sequenziamento di nuova generazione (NGS).
- Progetto NEXTGEN “Next Generation methods to preserve farm animal biodiversity by optimizing present and future breeding options”.  
*Periodo:* 2012–2014     *Finanziatore:* European Commission, 7th Framework Programme.  
Il contributo in questo progetto riguarda lo sviluppo, implementazione e sperimentazione di metodi algoritmici efficienti per la ricostruzione di aplotipi a partire da dati di sequenziamento di nuova generazione.
- Progetto PROZOO “Applicazione della genomica alla risoluzione di problemi di fertilità, resistenza alle malattie e assicurazione della qualità dei prodotti in bovini e suini”.  
*Periodo:* 2010–2014     *Finanziatori:* Fondazione Cariplo e Regione Lombardia.  
Il contributo in questo progetto riguarda lo sviluppo, implementazione e sperimentazione di metodi algoritmici efficienti per l’inferenza di aplotipi a partire da genotipi in grandi popolazioni animali.

## Borse di Studio e Assegni di Ricerca

- Vincitore di assegno di ricerca su bando competitivo con valutazione basata sul progetto di ricerca presentato dal candidato.  
Titolo: “Metodi e modelli algoritmici per l’assemblaggio e la correzione di dati di future-generation sequencing” (SSD: INF/01).  
Durata contratto: 1 gennaio 2016–31 dicembre 2017 (non goduto).  
Bando dell’Università degli Studi di Milano-Bicocca.
- Borsa di studio ministeriale per la frequenza del Dottorato di Ricerca (nov. 2006–ott. 2009).

- Borsa di studio per giovani promettenti sul tema “Sviluppo e Utilizzo di Tecniche di Soft Computing Applicate all’Analisi di Sequenze Biologiche” (mag.-ott. 2006). Responsabile: Dott. Giulio Pavesi, Università degli Studi di Milano.

## **Abilitazioni**

- Abilitazione Scientifica Nazionale per le funzioni di professore di seconda fascia per il settore 01/B1 - Informatica (dal 29/04/2021 al 29/04/2030).
- “Qualification aux fonctions de Maître de Conférences” per le sezioni 27 e 61 del CNU francese (rispettivamente equivalenti ai settori scientifici disciplinari INF/01 e ING-INF/05), campagna 2016.

## **Collaborazioni di Ricerca e Partecipazioni a Gruppi di Ricerca**

I miei interessi di ricerca si concentrano sullo studio della complessità computazionale di problemi di ottimizzazione e sul disegno e sperimentazione di algoritmi risolutivi efficienti utilizzando diverse tecniche, prevalentemente di carattere combinatorio.

Le mie attività di ricerca e le mie collaborazioni riguardano in particolar modo i seguenti temi.

**Indicizzazione di collezioni di testi e applicazioni in bioinformatica.** Collaborazioni con Prof. Kucherov (Université Paris-Est Marne-la-Vallée, Parigi) e con Dott. Giovanna Rosone (Univ. di Pisa). La collaborazione ha principalmente riguardato la progettazione di algoritmi efficienti in tempo e spazio basati su strutture di self-indexing per trattare dati massivi quali quelli prodotti dalle moderne tecniche di sequenziamento.

**Complessità e algoritmi parametrici.** Collaborazioni con Prof. Klaus (CWI, Amsterdam, ora Univ. di Düsseldorf, Germania), con Prof. Pisanti (Univ. di Pisa) e con Prof. Dondi (Univ. di Bergamo). La collaborazione si è focalizzata sullo studio della complessità computazionale (anche in senso parametrico) e sulla progettazione di algoritmi parametrici per problemi combinatori NP-hard con importanti applicazioni in genetica di popolazioni.

**Algoritmi per inferenza di aplotipi.** Collaborazione con Prof. Jiang (UC Riverside, USA) riguardante la definizione di problemi combinatori per l’inferenza di aplotipi, lo studio della loro complessità computazionale e lo sviluppo di algoritmi per la loro risoluzione. La sperimentazione su dati reali si è svolta in collaborazione con Parco Tecnologico Padano (Lodi) e con l’Istituto di Biologia e Biotecnologia Agraria del Consiglio Nazionale delle Ricerche nelle persone di Dott. Stella (direttrice scientifica PTP), Dott. Biscarini e Dott. Biffani (ricercatori IBBA-CNR).

**Ricostruzione vincolata e confronto vincolato di sequenze.** Collaborazioni con Prof. Beerenwinkel (ETH, Zurigo) e Prof. Dondi (Univ. di Bergamo). La collaborazione ha riguardato lo studio della complessità computazionale (anche in senso parametrico) di problemi di ottimizzazione riguardanti la ricostruzione di sequenze e copertura di grafi diretti aciclici (DAG) soggetti a vincoli (presenza di sottosequenze o di sottopercorsi nella soluzione) derivanti da applicazioni in genomica comparativa e assemblaggio di genomi virali.

**Algoritmi per l'assemblaggio e l'analisi di sequenze genomiche e trascrittomiche.** Collaborazioni con Prof. Pesole (Head of Node di ELIXIR-IIB il nodo italiano di ELIXIR, la principale infrastruttura europea per la bioinformatica), con Dott. Picardi (Univ. di Bari), con Prof. Alberghina e Dott. Chiaradonna (Univ. di Milano-Bicocca), con Dott. Castrignanò (CINECA SuperComputing Applications and Innovation Department, Roma), con Dott. Milanesi e Dott. Merelli (ricercatori ITB-CNR), con Dott. Montini e Dott. Beretta (San Raffaele Telethon Institute for Gene Therapy) e con Prof. De Felice, Dott. Zizza (Univ. di Salerno). In quest'area ho contribuito alla realizzazione del software principale usato per popolare il database ASPIC-DB relativo alla predizione della struttura dei geni, nonché allo sviluppo di metodi per l'analisi integrativa di dati genomici e trascrittomici.

**Misure di difficoltà in programmazione genetica.** Collaborazioni con Prof. Collard (Univ. di Nizza, Francia), con Prof. Verel (Univ. Littoral Côte d'Opale, Francia), con Prof. Tomassini (Univ. di Losanna, Svizzera) e con Prof. Vanneschi (NOVA IMS, Lisbona, Portogallo). La collaborazione ha riguardato la definizione e lo studio di misure per caratterizzare a priori la capacità di tecniche di GP nel trovare buone soluzioni a formalizzazioni di problemi di ottimizzazione.

Sono membro del Centro di Ricerca Interdipartimentale "*Bicocca Bioinformatics Biostatistics and Bioimaging centre*" – B4, Università degli Studi di Milano-Bicocca.

Ho collaborazioni stabili su diversi temi con i membri dei laboratori BIMIB (Bioinformatica e Calcolo Naturale) e BIAS (Bioinformatica e Algoritmica Sperimentale) dell'Università degli Studi di Milano-Bicocca (specificatamente Prof. Bonizzoni, Prof. Della Vedova e Prof. Rizzi).

## Prodotti della Ricerca

I contributi della mia attività scientifica sono stati presentati in 30 lavori a riviste di rilevanza internazionale, in 17 atti di convegni con peer-review e in 2 capitoli di libro.

Inoltre i risultati originali hanno contribuito alla realizzazione di 15 prodotti software open source resi disponibili alla comunità scientifica per l'utilizzo, il confronto e/o la replicazione delle sperimentazioni illustrate nelle pubblicazioni, tra i quali riporto i principali:

- **Pintron**, software per la predizione di strutture geniche, utilizzato per popolare ASPIC-DB, il database di eventi di splicing alternativo in geni umani [J7, J1].
- **Shark**, software per il filtraggio di dati di sequenziamento del trascrittoma, in grado di velocizzare analisi esistenti su insiemi di geni [J26].
- **Heu-MCHC** e **reHC-\***, due software per l'inferenza di aplotipi in popolazioni strutturate, rispettivamente in presenza di ricombinazioni e mutazioni [J5] e di informazioni mancanti e errori [J8].
- **LSG** e **FSG**, due software per la ricostruzione efficiente di *string graphs* mediante visita della BWT di una collezione di stringhe [J21, J20, J15].
- **bwt-lcp-em** e **bwt-lcp-parallel**, due algoritmi per la costruzione efficiente di BWT e array LCP di una collezione di stringhe in memoria esterna [J23] e multithread [J27], rispettivamente.
- **HapCol**, software che implementa un algoritmo parametrico per la ricostruzione di aplotipi da frammenti ottenuti da tecniche di sequenziamento di terza generazione [J14].  
Una sua estensione è stata inclusa in WhatsHap, uno dei tool utilizzati nel progetto Genome In A Bottle (GIAB) guidato dal National Institute of Standards and Technology (NIST) al fine di ottenere un genoma umano diploide di riferimento.

Membro del team di **BioConda**, uno dei gestori di pacchetti per la distribuzione di software scientifico più utilizzato in bioinformatica (circa 21k commit e 1k star su GitHub). Questo progetto offre una collezione di più di 3.000 pacchetti di software scientifici costantemente mantenuti dalla comunità al fine di garantire e supportare la riproducibilità di risultati di esperimenti e analisi computazionali [J22].

## Supervisione di Attività di Ricerca Post-Laurea

### Assegni di Ricerca

Responsabile scientifico dell'assegno di ricerca dal titolo "Algoritmi efficienti in pangenomica computazionale" vinto dal Dott. Jorge Avila Cartes, 12 mesi.

### Borse di Studio

- "Algoritmi efficienti per pangenomica computazionale comparativa", Dott. Simone Ciccolella, 6 mesi
- "Algoritmi per l'identificazione di eventi di splicing alternativo", Dott. Luca Denti, 8 mesi

### Dottorandi di Ricerca




Co-supervisione del dottorando di ricerca in Informatica Simone Ciccolella (titolo: "*Practical algorithms for Computational Phylogenetics*").

Co-supervisione di parte delle attività di ricerca svolte durante il corso di dottorato di ricerca in Informatica di Dott. Marco Previtali (titolo: "*Self-indexing for de novo assembly*") e Dott. Simone Zaccaria (titolo: "*Inferring Genomic Variants and their Evolution: Combinatorial Optimization for Haplotype Assembly and Quantification of Intra-Tumor Heterogeneity*").

## Elenco delle Pubblicazioni

Il simbolo  indica le pubblicazioni liberamente disponibili.

### Articoli su Riviste di Rilevanza Internazionale (con peer-review)

- [J30] Bonizzoni, P., Costantini, M., De Felice, C., Petescia, A., **Pirola, Y.**, Previtali, M., Rizzi, R., Stoye, J., Zaccagnino, R., and Zizza, R. "Numeric Lyndon-based feature embedding of sequencing reads for machine learning approaches." *Inf. Sci.* 607 (2022), 458–476. DOI: [10.1016/j.ins.2022.06.005](https://doi.org/10.1016/j.ins.2022.06.005).  
arXiv: [2202.13884v2](https://arxiv.org/abs/2202.13884v2) [q-bio.GN]. In: ISI WoS, Scopus IF (JCR 2020): 6.795. 
- [J29] Ciccolella, S., Denti, L., Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, and Previtali, M. "MALVIRUS: an integrated application for viral variant analysis." *BMC Bioinformatics* 22.15 (2022), 625.  
DOI: [10.1186/s12859-022-04668-0](https://doi.org/10.1186/s12859-022-04668-0). In: ISI WoS, Scopus IF (JCR 2020): 3.169. 
- [J28] Baaijens, J. A., Bonizzoni, P., Boucher, C., Della Vedova, G., **Pirola, Y.**, Rizzi, R., and Sirén, J. "Computational graph pangenomics: a tutorial on data structures and their applications." *Nat. Comput.* 21 (2022), 81–108. DOI: [10.1007/s11047-022-09882-6](https://doi.org/10.1007/s11047-022-09882-6). In: ISI WoS, Scopus IF (JCR 2020): 1.690. 

- [J27] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “Computing the multi-string BWT and LCP array in external memory.” *Theor. Comput. Sci.* 862 (2021), 42–58. DOI: [10.1016/j.tcs.2020.11.041](https://doi.org/10.1016/j.tcs.2020.11.041). In: ISI WoS, Scopus IF (JCR 2020): 0.827.
- [J26] Denti, L., **Pirola, Y.**, Previtali, M., Ceccato, T., Della Vedova, G., Rizzi, R., and Bonizzoni, P. “Shark: fishing relevant reads in an RNA-Seq sample.” *Bioinformatics* 37.4 (2021), 464–472. DOI: [10.1093/bioinformatics/btaa779](https://doi.org/10.1093/bioinformatics/btaa779). In: ISI WoS, Scopus IF (JCR 2020): 6.937.
- [J25] Rizzi, R., Beretta, S., Patterson, M., **Pirola, Y.**, Previtali, M., Della Vedova, G., and Bonizzoni, P. “Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era.” *Quant. Biol.* 7.4 (2019), 278–292. DOI: [10.1007/s40484-019-0181-x](https://doi.org/10.1007/s40484-019-0181-x). In: ISI WoS, Scopus JCI (2020): 0.35.
- [J24] Calabria, A., Beretta, S., Merelli, I., Spinozzi, G., Brasca, S., **Pirola, Y.**, Benedicenti, F., Tenderini, E., Bonizzoni, P., Milanese, L., and Montini, E. “ $\gamma$ -TRIS: a graph-algorithm for comprehensive identification of vector genomic insertion sites.” *Bioinformatics* 36.5 (2020), 1622–1624. DOI: [10.1093/bioinformatics/btz747](https://doi.org/10.1093/bioinformatics/btz747). In: ISI WoS, Scopus IF (JCR 2020): 6.937.
- [J23] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “Multithread Multistring Burrows–Wheeler Transform and Longest Common Prefix Array.” *J. Comput. Biol.* 26.9 (2019), 948–961. DOI: [10.1089/cmb.2018.0230](https://doi.org/10.1089/cmb.2018.0230). In: ISI WoS, Scopus IF (JCR 2020): 1.479.
- [J22] Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., and The Bioconda Team (including **Pirola, Y.**) “Bioconda: sustainable and comprehensive software distribution for the life sciences.” *Nature Methods* 15.7 (2018), 475–476. DOI: [10.1038/s41592-018-0046-7](https://doi.org/10.1038/s41592-018-0046-7). In: ISI WoS, Scopus IF (JCR 2020): 28.547.
- [J21] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “FSG: Fast String Graph Construction for De Novo Assembly.” *J. Comput. Biol.* 24.10 (2017), 953–968. DOI: [10.1089/cmb.2017.0089](https://doi.org/10.1089/cmb.2017.0089). In: ISI WoS, Scopus IF (JCR 2020): 1.479.
- [J20] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “An External-Memory Algorithm for String Graph Construction.” *Algorithmica* 78.2 (2017), 394–424. DOI: [10.1007/s00453-016-0165-4](https://doi.org/10.1007/s00453-016-0165-4). In: ISI WoS, Scopus IF (JCR 2020): 0.791.
- [J19] Biscarini, F., Schwarzenbacher, H., Pausch, H., Nicolazzi, E. L., **Pirola, Y.**, and Biffani, S. “Use of SNP genotypes to identify carriers of harmful recessive mutations in cattle populations.” *BMC Genomics* 17 (2016), 857. DOI: [10.1186/s12864-016-3218-9](https://doi.org/10.1186/s12864-016-3218-9). In: ISI WoS, Scopus IF (JCR 2020): 3.969.
- [J18] Chiaradonna, F., **Pirola, Y.**, Ricciardiello, F., and Palorini, R. “Transcriptional profiling of immortalized and K-ras-transformed mouse fibroblasts upon PKA stimulation by forskolin in low glucose availability.” *Genomics Data* 9 (2016), 100–104. DOI: [10.1016/j.gdata.2016.07.004](https://doi.org/10.1016/j.gdata.2016.07.004). In: ISI WoS, Scopus.
- [J17] Bonizzoni, P., Dondi, R., Klau, G. W., **Pirola, Y.**, Pisanti, N., and Zaccaria, S. “On the Minimum Error Correction Problem for Haplotype Assembly in Diploid and Polyploid Genomes.” *J. Comput. Biol.* 23.9 (2016), 718–736. DOI: [10.1089/cmb.2015.0220](https://doi.org/10.1089/cmb.2015.0220). In: ISI WoS, Scopus IF (JCR 2020): 1.479.
- [J16] Palorini, R., Votta, G., **Pirola, Y.**, De Vitto, H., De Palma, S., Airolidi, C., Vasso, M., Ricciardiello, F., Lombardi, P. P., Cirulli, C., Rizzi, R., Nicotra, F., Hiller, K., Gelfi, C., Alberghina, L., and Chiaradonna, F. “Protein Kinase A Activation Promotes Cancer Cell Resistance to Glucose Starvation and Anoikis.” *PLoS Genet.* 12.3 (2016), 1–41. DOI: [10.1371/journal.pgen.1005931](https://doi.org/10.1371/journal.pgen.1005931). In: ISI WoS, Scopus IF (JCR 2020): 5.917.
- [J15] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “LSG: An External-Memory Tool to Compute String Graphs for NGS Data Assembly.” *J. Comput. Biol.* 23.3 (2016), 137–149. DOI: [10.1089/cmb.2015.0172](https://doi.org/10.1089/cmb.2015.0172). In: ISI WoS, Scopus IF (JCR 2020): 1.479.

- [J14] **Pirola, Y.**, Zaccaria, S., Dondi, R., Klau, G. W., Pisanti, N., and Bonizzoni, P. “HapCol: Accurate and Memory-Efficient Haplotype Assembly from Long Reads.” *Bioinformatics* 32.11 (2016), 1610–1617. DOI: [10.1093/bioinformatics/btv495](https://doi.org/10.1093/bioinformatics/btv495). In: ISI WoS, Scopus IF (JCR 2020): 6.937.
- [J13] Beerenwinkel, N., Beretta, S., Bonizzoni, P., Dondi, R., and **Pirola, Y.** “Covering Pairs in Directed Acyclic Graphs.” *Comput. J.* 58.7 (2015), 1673–1686. DOI: [10.1093/comjnl/bxu116](https://doi.org/10.1093/comjnl/bxu116). In: ISI WoS, Scopus IF (JCR 2020): 1.494.
- [J12] Batini, C., Bonizzoni, P., Comerio, M., Dondi, R., **Pirola, Y.**, and Salandra, F. “A Clustering Algorithm for Planning the Integration Process of a Large Number of Conceptual Schemas.” *J. Comput. Sci. Technol.* 30.1 (2015), 214–224. DOI: [10.1007/s11390-015-1514-5](https://doi.org/10.1007/s11390-015-1514-5). In: ISI WoS, Scopus IF (JCR 2020): 1.571.
- [J11] Beretta, S., Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, and Rizzi, R. “Modeling Alternative Splicing Variants from RNA-Seq Data with Isoform Graphs.” *J. Comput. Biol.* 21.1 (2014), 16–40. DOI: [10.1089/cmb.2013.0112](https://doi.org/10.1089/cmb.2013.0112). In: ISI WoS, Scopus IF (JCR 2020): 1.479.
- [J10] Bonizzoni, P., Della Vedova, G., Dondi, R., and **Pirola, Y.** “Parameterized Complexity of  $k$ -Anonymity: Hardness and Tractability.” *J. Comb. Optim.* 26.1 (2013), 19–43. DOI: [10.1007/s10878-011-9428-9](https://doi.org/10.1007/s10878-011-9428-9). In: ISI WoS, Scopus IF (JCR 2020): 1.195.
- [J9] Bonizzoni, P., Dondi, R., and **Pirola, Y.** “Maximum Disjoint Paths on Edge-Colored Graphs: Approximability and Tractability.” *Algorithms* 6.1 (2013), 1–11. DOI: [10.3390/a6010001](https://doi.org/10.3390/a6010001). In: ISI WoS, Scopus JCI (2020): 0.48.
- [J8] **Pirola, Y.**, Della Vedova, G., Biffani, S., Stella, A., and Bonizzoni, P. “A Fast and Practical Approach to Genotype Phasing and Imputation on a Pedigree with Erroneous and Incomplete Information.” *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9.6 (2012), 1582–1594. DOI: [10.1109/TCBB.2012.100](https://doi.org/10.1109/TCBB.2012.100). In: ISI WoS, Scopus IF (JCR 2020): 3.710.
- [J7] **Pirola, Y.**, Rizzi, R., Picardi, E., Pesole, G., Della Vedova, G., and Bonizzoni, P. “Pintron: A Fast Method for Detecting the Gene Structure Due to Alternative Splicing Via Maximal Pairings of a Pattern and a Text.” *BMC Bioinformatics* 13.S5 (2012), S2. DOI: [10.1186/1471-2105-13-S5-S2](https://doi.org/10.1186/1471-2105-13-S5-S2). In: ISI WoS, Scopus IF (JCR 2020): 3.169.
- [J6] Vanneschi, L., **Pirola, Y.**, Mauri, G., Tomassini, M., Collard, P., and Verel, S. “A Study of Neutrality of Boolean Function Landscapes in Genetic Programming.” *Theor. Comput. Sci.* 425 (2012), 34–57. DOI: [10.1016/j.tcs.2011.03.011](https://doi.org/10.1016/j.tcs.2011.03.011). In: ISI WoS, Scopus IF (JCR 2020): 0.827.
- [J5] **Pirola, Y.**, Bonizzoni, P., and Jiang, T. “An Efficient Algorithm for Haplotype Inference on Pedigrees with Recombinations and Mutations.” *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9.1 (2012), 12–25. DOI: [10.1109/TCBB.2011.51](https://doi.org/10.1109/TCBB.2011.51). In: ISI WoS, Scopus IF (JCR 2020): 3.710.
- [J4] Bonizzoni, P., Della Vedova, G., Dondi, R., and **Pirola, Y.** “Variants of Constrained Longest Common Subsequence.” *Inf. Process. Lett.* 110.20 (2010), 877–881. DOI: [10.1016/j.ipl.2010.07.015](https://doi.org/10.1016/j.ipl.2010.07.015). In: ISI WoS, Scopus IF (JCR 2020): 0.959.
- [J3] Bonizzoni, P., Della Vedova, G., Dondi, R., **Pirola, Y.**, and Rizzi, R. “Pure Parsimony Xor Haplotyping.” *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7.4 (2010), 598–610. DOI: [10.1109/TCBB.2010.52](https://doi.org/10.1109/TCBB.2010.52). In: ISI WoS, Scopus IF (JCR 2020): 3.710.
- [J2] Della Vedova, G., Dondi, R., Jiang, T., Pavesi, G., **Pirola, Y.**, and Wang, L. “Beyond Evolutionary Trees.” *Nat. Comput.* 9.2 (2010), 421–435. DOI: [10.1007/s11047-009-9156-6](https://doi.org/10.1007/s11047-009-9156-6). In: ISI WoS, Scopus IF (JCR 2020): 1.690.



- [J1] Bonizzoni, P., Mauri, G., Pesole, G., Picardi, E., **Pirola, Y.**, and Rizzi, R. “Detecting Alternative Gene Structures from Spliced ESTs: A Computational Approach.” *J. Comput. Biol.* 16.1 (2009), 43–66. DOI: [10.1089/cmb.2008.0028](https://doi.org/10.1089/cmb.2008.0028). In: ISI WoS, Scopus IF (JCR 2020): 1.479.

### Atti di Convegni di Rilevanza Internazionale (con peer-review)

- [C17] Bonizzoni, P., Petescia, A., **Pirola, Y.**, Rizzi, R., Zaccagnino, R., and Zizza, R. “KFinger: Capturing Overlaps Between Long Reads by Using Lyndon Fingerprints.” In: *Bioinformatics and Biomedical Engineering (IWBBIO)*. Vol. 13347. LNCS. Springer, 2022, 3–12. DOI: [10.1007/978-3-031-07802-6\\_37](https://doi.org/10.1007/978-3-031-07802-6_37). In: ISI WoS, Scopus.
- [C16] Bonizzoni, P., De Felice, C., **Pirola, Y.**, Rizzi, R., Zaccagnino, R., and Zizza, R. “Can Formal Languages Help Pangenomics to Represent and Analyze Multiple Genomes?” In: *Developments in Language Theory (DLT)*. Vol. 13257. LNCS. Springer, 2022, 3–12. DOI: [10.1007/978-3-031-05578-2\\_1](https://doi.org/10.1007/978-3-031-05578-2_1). In: ISI WoS, Scopus.
- [C15] Bonizzoni, P., De Felice, C., Petescia, A., **Pirola, Y.**, Rizzi, R., Stoye, J., Zaccagnino, R., and Zizza, R. “Can We Replace Reads by Numeric Signatures? Lyndon Fingerprints as Representations of Sequencing Reads for Machine Learning.” In: *Algorithms for Computational Biology (ALCoB)*. Vol. 12715. LNCS. Springer, 2021, 16–28. DOI: [10.1007/978-3-030-74432-8\\_2](https://doi.org/10.1007/978-3-030-74432-8_2). In: ISI WoS, Scopus.
- [C14] Bonizzoni, P., Della Vedova, G., Nicosia, S., **Pirola, Y.**, Previtali, M., and Rizzi, R. “Divide and Conquer Computation of the Multi-string BWT and LCP Array.” In: *Computability in Europe (CiE)*. Vol. 10936. LNCS. Springer, 2018, 107–117. DOI: [10.1007/978-3-319-94418-0\\_11](https://doi.org/10.1007/978-3-319-94418-0_11). In: ISI WoS, Scopus.
- [C13] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “FSG: Fast String Graph Construction for De Novo Assembly of reads data.” In: *Bioinformatics Research and Applications (ISBRA)*. Vol. 9683. LNCS. Springer, 2016, 27–39. DOI: [10.1007/978-3-319-38782-6\\_3](https://doi.org/10.1007/978-3-319-38782-6_3). In: ISI WoS, Scopus.
- [C12] Bonizzoni, P., Dondi, R., Klau, G. W., **Pirola, Y.**, Pisanti, N., and Zaccaria, S. “On the Fixed Parameter Tractability and Approximability of the Minimum Error Correction problem.” In: *Combinatorial Pattern Matching (CPM)*. Vol. 9133. LNCS. Springer, 2015, 100–113. DOI: [10.1007/978-3-319-19929-0\\_9](https://doi.org/10.1007/978-3-319-19929-0_9). In: Inspec, Scopus.
- [C11] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, Previtali, M., and Rizzi, R. “Constructing String Graphs in External Memory.” In: *Algorithms in Bioinformatics (WABI)*. Vol. 8701. LNCS. Springer, 2014, 311–325. DOI: [10.1007/978-3-662-44753-6\\_23](https://doi.org/10.1007/978-3-662-44753-6_23). In: ISI WoS, Scopus.
- [C10] Beerenwinkel, N., Beretta, S., Bonizzoni, P., Dondi, R., and **Pirola, Y.** “Covering Pairs in Directed Acyclic Graphs.” In: *Language and Automata Theory and Applications (LATA)*. Vol. 8370. LNCS. Springer, 2014, 126–137. DOI: [10.1007/978-3-319-04921-2\\_10](https://doi.org/10.1007/978-3-319-04921-2_10). In: ISI WoS, Scopus.
- [C9] **Pirola, Y.**, Della Vedova, G., Bonizzoni, P., Stella, A., and Biscarini, F. “Haplotype-based prediction of gene alleles using pedigrees and SNP genotypes.” In: *Bioinformatics, Computational Biology, and Biomedical Informatics (ACM BCB)*. ACM, 2013, 33–41. DOI: [10.1145/2506583.2506592](https://doi.org/10.1145/2506583.2506592). In: Scopus.
- [C8] **Pirola, Y.**, Della Vedova, G., Biffani, S., Stella, A., and Bonizzoni, P. “A fast and practical approach to genotype phasing and imputation on a pedigree with erroneous and incomplete information.” In: *Computational Advances in Bio and medical Sciences (ICCABS)*. IEEE, 2012. DOI: [10.1109/ICCABS.2012.6182643](https://doi.org/10.1109/ICCABS.2012.6182643). In: Inspec, Scopus.
- [C7] Bonizzoni, P., Della Vedova, G., **Pirola, Y.**, and Rizzi, R. “PItron: a fast method for gene structure prediction via maximal pairings of a pattern and a text.” In: *Computational Advances in Bio and medical Sciences (ICCABS)*. IEEE, 2011, 33–39. DOI: [10.1109/ICCABS.2011.5729935](https://doi.org/10.1109/ICCABS.2011.5729935). In: Inspec, Scopus.




- [C6] **Pirola, Y.**, Bonizzoni, P., and Jiang, T. “Haplotype Inference on Pedigrees with Recombinations and Mutations.” In: *Algorithms in Bioinformatics (WABI)*. Vol. 6293. LNCS. Springer, 2010, 148–161. DOI: [10.1007/978-3-642-15294-8\\_13](https://doi.org/10.1007/978-3-642-15294-8_13). In: ISI WoS, Scopus.
- [C5] Bonizzoni, P., Della Vedova, G., Dondi, R., and **Pirola, Y.** “Parameterized Complexity of k-Anonymity: Hardness and Tractability.” In: *Combinatorial Algorithms (IWOCA)*. Vol. 6460. LNCS. Springer, 2011, 242–255. DOI: [10.1007/978-3-642-19222-7\\_25](https://doi.org/10.1007/978-3-642-19222-7_25). In: ISI WoS, Scopus.
- [C4] Bonizzoni, P., Della Vedova, G., Dondi, R., **Pirola, Y.**, and Rizzi, R. “Minimum Factorization Agreement of Spliced ESTs.” In: *Algorithms in Bioinformatics (WABI)*. Vol. 5724. LNCS. Springer, 2009, 1–12. DOI: [10.1007/978-3-642-04241-6\\_1](https://doi.org/10.1007/978-3-642-04241-6_1). In: ISI WoS, Scopus.
- [C3] Bonizzoni, P., Della Vedova, G., Dondi, R., **Pirola, Y.**, and Rizzi, R. “Pure Parsimony Xor Haplotyping.” In: *Bioinformatics Research and Applications (ISBRA)*. Vol. 5542. LNCS. Springer, 2009, 186–197. DOI: [10.1007/978-3-642-01551-9\\_19](https://doi.org/10.1007/978-3-642-01551-9_19). In: ISI WoS, Scopus.
- [C2] Vanneschi, L., Tomassini, M., Collard, P., Verel, S., **Pirola, Y.**, and Mauri, G. “A Comprehensive View of Fitness Landscapes with Neutrality and Fitness Clouds.” In: *Genetic Programming (EuroGP)*. Vol. 4445. LNCS. Springer, 2007, 241–250. DOI: [10.1007/978-3-540-71605-1\\_22](https://doi.org/10.1007/978-3-540-71605-1_22). In: ISI WoS, Scopus.
- [C1] Vanneschi, L., **Pirola, Y.**, and Collard, P. “A quantitative study of neutrality in GP boolean landscapes.” In: *Genetic and Evolutionary Computation (GECCO)*. ACM, 2006, 895–902. DOI: [10.1145/1143997.1144152](https://doi.org/10.1145/1143997.1144152). In: ISI WoS, Scopus.

### Capitoli di libro

- [B2] Dondi, R. and **Pirola, Y.** “Beyond Evolutionary Trees.” In: *Encyclopedia of Algorithms*. Ed. by M.-Y. Kao. Springer, 2016, 183–189. ISBN: 978-3-642-27848-8. DOI: [10.1007/978-3-642-27848-8\\_599-1](https://doi.org/10.1007/978-3-642-27848-8_599-1).
- [B1] Bonizzoni, P., Della Vedova, G., Pesole, G., Picardi, E., **Pirola, Y.**, and Rizzi, R. “Transcriptome Assembly and Alternative Splicing Analysis.” In: *RNA Bioinformatics*. Ed. by E. Picardi. Vol. 1269. Methods in Molecular Biology. Springer, 2015, 173–188. ISBN: 978-1-4939-2290-1. DOI: [10.1007/978-1-4939-2291-8\\_11](https://doi.org/10.1007/978-1-4939-2291-8_11). In: ISI WoS, Scopus.

### Tesi di Dottorato di Ricerca

- [T1] **Pirola, Y.** “Combinatorial Problems in Studies of Genetic Variations: Haplotyping and Transcript Analysis.” PhD thesis. Università degli Studi di Milano-Bicocca, 2010. HDL: [10281/7891](https://hdl.handle.net/10281/7891). 

## Eventi e Convegni

### Seminari su Invito

- Università degli Studi di Milano su invito di Prof. Giovanni Righini, 8 Febbraio 2010.
- Parco Tecnologico Padano (Lodi) su invito di Dott. Alessandra Stella, 25 Marzo 2010.

### Partecipazione come Relatore a Convegni Scientifici

- *9th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, Giugno 2022, Gran Canaria, Spagna (online), in qualità di relatore del lavoro [C17].
- *15th Bioinformatics and Computational Biology Conference (BBCC)*, Novembre 2020, Napoli (online), in qualità di relatore del lavoro “MALVIRUS: an integrated web application for viral variant calling”. DOI: [10.7490/F1000RESEARCH.1118377.1](https://doi.org/10.7490/F1000RESEARCH.1118377.1).
- *21th Bioinformatics Open Source Conference (BOSC)*, Luglio 2020, Toronto, Canada (online), in qualità di relatore del lavoro “MALVIRUS: viral variant calling made easy”. DOI: [10.1101/2020.05.05.076992](https://doi.org/10.1101/2020.05.05.076992).
- *6th Int. Workshop “Data Structures in Bioinformatics”*, Febbraio 2020, Rennes, Francia, in qualità di relatore del lavoro “Shark: Fishing in a sample to discard irrelevant RNA-Seq reads”. DOI: [10.1101/836130](https://doi.org/10.1101/836130).
- *8th Int. Conf. on Language and Automata Theory and Applications (LATA)*, Marzo 2014, Madrid, Spain, in qualità di relatore del lavoro [C10].
- *Workshop “Combinatorial structures for sequence analysis in bioinformatics”*, Novembre 2013, Milano, Italia, in qualità di relatore del lavoro “Combinatorial structures and NGS data in transcriptomics: some results”.
- *4th ACM Int. Conf. on Bioinformatics, Computational Biology, and Biomedical Informatics (ACM BCB)*, Settembre 2013, Washington DC, USA, in qualità di relatore del lavoro [C9].
- *14th Italian Conf. on Theoretical Computer Science (ICTCS)*, Settembre 2013, Palermo, Italia, in qualità di relatore del lavoro “Covering pairs in directed acyclic graphs” (peer-reviewed).
- *IEEE Int. Conf. on Computational Advances in Bio and medical Sciences (ICCABS)*, Febbraio 2012, Las Vegas NV, USA, in qualità di relatore del lavoro [C8].
- *IEEE Int. Conf. on Computational Advances in Bio and medical Sciences (ICCABS)*, Febbraio 2011, Orlando FL, USA, in qualità di relatore del lavoro [C7].
- *Int. Workshop on Algorithms in Bioinformatics (WABI)*, Settembre 2010, Liverpool, UK, in qualità di relatore del lavoro [C6].
- *Int. Workshop on Algorithms in Bioinformatics (WABI)*, Settembre 2009, Philadelphia PA, USA, in qualità di relatore del lavoro [C4].
- *Int. Symp. on Bioinformatics Research and Applications (ISBRA)*, Maggio 2009, Ft. Lauderdale FL, USA, in qualità di relatore del lavoro [C3].

### Visite di Ricerca

- Visita di ricerca presso il Centrum Wiskunde & Informatica, Amsterdam, Olanda, dal 14 al 17 luglio 2015, per la collaborazione con il Dott. Gunnar Klau (leader of Life Sciences group) e il Prof. Alexander Schönhuth sulla progettazione di algoritmi parametrici per problemi di genetica di popolazione.

- Visita di ricerca presso l'Université Paris-Est Marne-la-Vallée, Parigi, Francia, dal 12 al 18 gennaio 2014, per la collaborazione con il Dott. Gregory Kucherov sui temi del progetto PRIN 2010/11.
- *Visiting Scholar* presso la University of California, Riverside, USA, da Febbraio 2009 a Giugno 2009, invitato da Prof. Tao Jiang.

## Servizio alla Comunità Scientifica

Co-organizer (con S. Pissis, CWI) di “*Computational Pangenomics: Algorithms & Applications*”, workshop satellite di “*ECCB 2020, 19th European Conference on Computational Biology*”, Settembre 2020, Barcelona, Spain.

Local Technical Coordinator (LTec) per l'Univ. degli Studi di Milano-Bicocca di ELIXIR-IIB, il nodo italiano di ELIXIR, la principale infrastruttura europea per la bioinformatica.

Membro del Programme Committee dei seguenti convegni internazionali:

- “ICTCS 2022, 23rd Italian Conf. on Theoretical Computer Science”, Settembre 2022, Roma, Italia
- “BBC 2022, 15th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2022, 22th Int. Conf. on Computational Science”, Giugno 2022, London, UK
- “TWBBIO 2022, 8th Int. Work-Conference on Bioinformatics and Biomedical Engineering”, Giugno 2022, Gran Canaria, Spain
- “BICOB 2022, 14th Int. Conf. on Bioinformatics and Computational Biology”, Marzo 2022, On-line
- “BBC 2021, 14th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2021, 21th Int. Conf. on Computational Science”, Giugno 2021, Kraków, Poland
- “HPC4COVID-19, High Performance Computing Methods and Interdisciplinary Applications for Fighting the COVID-19 Pandemic” parte di “IEEE BIBM 2020, 14th Int. Conf. on Bioinformatics and Biomedicine”, Dicembre 2020, Seoul, South Korea
- “BBC 2020, 13th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2020, 20th Int. Conf. on Computational Science”, Giugno 2020, Amsterdam, The Netherlands
- “TWBBIO 2020, 8th Int. Work-Conference on Bioinformatics and Biomedical Engineering”, Maggio 2020, Granada, Spain
- “BICOB 2020, 12th Int. Conf. on Bioinformatics and Computational Biology”, Marzo 2020, San Francisco CA, USA
- “BBC 2019, 12th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2019, 19th Int. Conf. on Computational Science”, Giugno 2019, Faro, Portugal
- “BBC 2018, 11th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2018, 18th Int. Conf. on Computational Science”, Giugno 2018, Wuxi, China
- “BBC 2017, 10th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2017, 17th Int. Conf. on Computational Science”, Giugno 2017, Zurigo, Svizzera
- “PDP 2017, 25th Euromicro Int. Conf. on Parallel, Distributed, and Network-Based Processing (Track: Advances in High-Performance Bioinformatics, Systems and Synthetic Biology)”, Marzo 2017, St. Petersburg, Russia

- “BBC 2016, 9th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2016, 16th Int. Conf. on Computational Science”, Giugno 2016, San Diego CA, USA
- “PDP 2016, 24th Euromicro Int. Conf. on Parallel, Distributed, and Network-Based Processing (Special Session on Advances in High-Performance Bioinformatics, Systems and Synthetic Biology)”, Febbraio 2016, Crete, Greece
- “BBC 2015, 8th Workshop on Biomedical and Bioinformatics Challenges for Computer Science” parte di “ICCS 2015, 15th Int. Conf. on Computational Science”, Giugno 2015, Reykjavík, Iceland
- “PDP 2015, 23rd Euromicro Int. Conf. on Parallel, Distributed, and Network-Based Processing (Special Session on Advances in High-Performance Bioinformatics, Systems and Synthetic Biology)”, Marzo 2015, Turku, Finland

Membro del Comitato Organizzativo dei seguenti convegni:

- Workshop “Data Structures in Bioinformatics 2021 (DSB)”, Febbraio 2021, Univ. degli Studi di Milano-Bicocca
- Conferenza “Computability in Europe 2013 (CiE)”, Luglio 2013, Univ. degli Studi di Milano-Bicocca
- Colloquium “Unconventional Models of Computation”, Settembre 2009, Politecnico di Milano
- “ASWorkshop - Alternative Splicing in Animals and Plants”, Ottobre 2008, Univ. degli Studi di Milano-Bicocca

Membro del Comitato Organizzativo della Scuola di Dottorato “*Introduction to Pangenomics*”, 4–8 Luglio 2022, Lake Como School of Advanced Studies.

Review editor di “Frontiers in Bioinformatics”, diretto da Prof. Adam Godzik (UC Riverside, CA, US). L’incarico di review editor comprende la responsabilità di gestire il processo di review dell’articolo e il nome dell’editore è pubblicamente indicato in ciascun articolo che ha gestito.

Membro del Reviewer Board di “Algorithms” (indicizzato da Web of Science e Scopus).

Reviewer per le seguenti riviste:

- “Discrete Applied Mathematics”
- “Bioinformatics”
- “Briefings in Bioinformatics”
- “IEEE/ACM Transactions on Computational Biology and Bioinformatics”
- “IEEE Access”
- “GigaScience”
- “Journal of Computational Science”
- “Journal of Biomedical Informatics”
- “Computers & Operations Research”
- “International Transactions in Operational Research”
- “Applied Sciences”
- “Mathematics”
- “Genes”

- “Processes”
- “Systems”
- “Technologies”
- “Concurrency and Computation: Practice and Experience”
- “BMC Genomics”
- “International Journal of Molecular Sciences”
- “International Journal of Bioinformatics Research and Applications”
- “BioMed Research International”

Reviewer per i seguenti convegni internazionali:

- BIBM 2022, IEEE Int. Conf. on Bioinformatics and Biomedicine, Las Vegas NV, USA
- WABI 2022, Workshop on Algorithms in Bioinformatics, Potsdam, Germania
- BIBM 2021, IEEE Int. Conf. on Bioinformatics and Biomedicine, Houston TX, USA
- ISBRA 2021, Int. Symp. on Bioinformatics Research and Applications, Shenzhen, China
- RECOMB-CG 2021, RECOMB International Satellite Workshop on Comparative Genomics, Singapore
- WABI 2021, Workshop on Algorithms in Bioinformatics, Gainesville FL, USA
- BIBM 2020, IEEE Int. Conf. on Bioinformatics and Biomedicine, Seoul, Korea
- APBC 2020, Asia Pacific Bioinformatics Conf., Seoul, Korea
- CIBB 2019, Int. Conf. on Computational Intelligence methods for Bioinformatics and Biostatistics, Bergamo, Italia
- SODA 2018, ACM-SIAM Symp. on Discrete Algorithms, New Orleans LA, USA
- IWOCA 2016, Int. Workshop on Combinatorial Algorithms, Helsinki, Finland
- AAIM 2016, Int. Conf. on Algorithmic Aspects in Information and Management, Bergamo, Italia
- WALCOM 2016, Int. Workshop on Algorithms and Computation, Kathmandu, Nepal
- IWOCA 2015, Int. Workshop on Combinatorial Algorithms, Verona, Italia
- ISBRA 2015, Int. Symp. on Bioinformatics Research and Applications, Norfolk VA, USA
- WABI 2014, Workshop on Algorithms in Bioinformatics, Wrocław, Polonia
- CIBB 2014, Int. Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Cambridge, UK
- ISBRA 2014, Int. Symp. on Bioinformatics Research and Applications, Zhangjiajie, China
- BIBM 2013, IEEE Int. Conf. on Bioinformatics and Biomedicine, Shanghai, China
- BSB 2013, Brazilian Symposium on Bioinformatics, Recife, Brasile
- CiE 2013, Computability in Europe, Milano, Italia
- CATS 2013, Computing: the Australasian Theory Symposium, Melbourne, Australia
- BIBM 2012, IEEE Int. Conf. on Bioinformatics and Biomedicine, Philadelphia PA, USA
- WABI 2012, Workshop on Algorithms in Bioinformatics, Ljubljana, Slovenia

- ECCB 2012, European Conf. on Computational Biology, Basel, Svizzera
- ICCABS 2011, IEEE Int. Conf. on Computational Advances in Bio and medical Sciences, Orlando FL, USA
- BIBM 2009, IEEE Int. Conf. on Bioinformatics and Biomedicine, Washington DC, USA

## **Supervisione di Stage e Tesi**

### **Tesi di Laurea Magistrale**

Co-supervisione delle seguenti tesi di Laurea Magistrale in Informatica conseguite presso l'Univ. degli Studi di Milano-Bicocca:

- "Implementazione di grafi di de Bruijn dinamici attraverso learned index"  
Studente: Riccardo Nigrelli, A.A. 2020/21.
- "Applicazioni della BWT posizionale e grafica in pangenomica computazionale"  
Studente: Gabriele Molteni, A.A. 2019/20.
- "Metodi efficienti per la classificazione di dati di metagenomica da NGS"  
Studente: Matteo Fumagalli, A.A. 2019/20.
- "Indel Reversal Distance"  
Studente: Simone Zaccaria, A.A. 2012/13.
- "Metodi di Consenso di Alberi di Suffisso nella Ricerca di Sottosequenze Comuni"  
Studente: Marcello Varisco, A.A. 2007/08.

### **Stage**

Co-supervisione dei seguenti stage curriculari per il conseguimento della Laurea in Informatica presso l'Univ. degli Studi di Milano-Bicocca:

- "Ottimizzazione di un software per la genotipizzazione di varianti"  
Studente: Marco Burgio, A.A. 2020/21.
- "Estensione di Algoritmi di Ricerca (pBWT) su Matrici di Aplotipi con GAP"  
Studente: Mattia Sgro, A.A. 2019/20.
- "Convertitore di formati di input per Celluloid"  
Studente: Danilo Fumagalli, A.A. 2019/20.
- "Grafici di pangenomi di riferimento con pygfa"  
Studente: Francesco Lapi, A.A. 2019/20.
- "Confronto di strutture dati di indicizzazione per il filtraggio di dati trascrittomici"  
Studente: Davide Pizzoli, A.A. 2019/20.
- "Inferenza di Aplotipi in Pedigree Multiallelici Tramite Risolutori SAT"  
Studente: Simone Zaccaria, A.A. 2010/11.
- "Algoritmi di Allineamento Basati su Alberi di Suffisso"  
Studente: Alberto Villa, A.A. 2007/08.

- “Metodi Combinatori di Predizione di Eventi di Splicing Associati a Polimorfismi”  
Studente: Maurizio Rorato, A.A. 2006/07.
- “Algoritmi per la Ricostruzione di Filogenesi da Matrici di Genotipi”  
Studente: Federica Musitelli, A.A. 2006/07.

## **Attività Didattiche**

### **Dottorato di Ricerca in Informatica**

Membro del collegio docenti del programma di Dottorato di Ricerca in Informatica (dal ciclo XXXVII) presso l’Univ. degli Studi di Milano-Bicocca.

- Insegnamento “Do we need data structures?”, A.A. 2019/20.  
Lezioni frontali, 4 ore.  
Insegnamento congiunto delle Scuole di Dottorato di Ricerca in Informatica dell’Univ. degli Studi di Milano e dell’Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Population-Based Optimisation Methods”, A.A. 2021/22.  
Lezioni frontali, 4 ore.  
Scuola di Dottorato di Ricerca in Informatica dell’Univ. degli Studi di Milano-Bicocca.

### **Master**

- Moduli “Genome assembling” e “Beyond genome assembly”, A.A. 2020/21.  
Lezioni frontali, 20 ore.  
Master di 2o livello “qOmics: quantitative methods for Omics Data”,  
Univ. degli Studi di Milano-Bicocca.

### **Scuole di Specializzazione di Area Medica e Sanitaria**

- Insegnamento “Abilità Informatiche”, A.A. 2018/19–2021/22.  
Lezioni frontali, 16 ore.  
Scuole di Specializzazione del Dip. di Medicina e Chirurgia, Univ. degli Studi di Milano-Bicocca.

### **Corsi di Laurea e Laurea Magistrale**

- Insegnamento “Algoritmi e Strutture Dati”, A.A. 2022/23.  
Lezioni frontali, 32 ore (4 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Algoritmi e Strutture Dati”, A.A. 2022/23.  
Esercitazioni frontali, 20 ore (2 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Algoritmi e Strutture Dati”, A.A. 2022/23.  
Esercitazioni di laboratorio, 40 ore (4 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.



- Modulo “Teoria della Computazione”, Insegnamento “Modelli e Computazione”, A.A. 2022/23.  
Lezioni frontali, 16 ore (2 cfu).  
Corso di Laurea Magistrale in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Laboratorio di Progettazione”, A.A. 2019/20–2022/23.  
Esercitazioni frontali, 12 ore (1 cfu).  
Corso di Laurea Magistrale in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Algoritmi e Strutture Dati”, A.A. 2019/20–2021/22.  
Esercitazioni frontali, 20 ore (2 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Algoritmi e Strutture Dati”, A.A. 2019/20–2021/22.  
Esercitazioni di laboratorio, 20 ore (2 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Elementi di Bioinformatica”, A.A. 2013/14.  
Esercitazioni di laboratorio, 16 ore (2 cfu).  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Modulo “Algoritmi e Strutture Dati 2”, Insegnamento “Algoritmi e Ricerca Operativa”, A.A. 2010/11.  
Esercitazioni frontali, 12 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Informatica Generale 2”, A.A. 2007/08–2008/09.  
Lezioni frontali, 30 ore.  
Corso di Laurea in Lingue e Letterature Straniere, Univ. degli Studi di Bergamo.
- Modulo “Analisi di Algoritmi”, Insegnamento “Tecniche di Analisi e Verifica”, A.A. 2007/08.  
Esercitazioni frontali, 12 ore.  
Corso di Laurea Magistrale in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Bioinformatica”, A.A. 2007/08.  
Esercitazioni frontali, 12 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Laboratorio di Linguaggi di Programmazione”, A.A. 2006/07.  
Esercitazioni di laboratorio, 24 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Laboratorio di Algoritmi e Strutture Dati”, A.A. 2005/06.  
Esercitazioni di laboratorio, 12 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Sistemi Operativi (Elementi e Complementi)”, A.A. 2005/06.  
Esercitazioni di laboratorio, 24 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.
- Insegnamento “Sistemi Operativi (Elementi e Complementi)”, A.A. 2004/05.  
Esercitazioni di laboratorio, 48 ore.  
Corso di Laurea in Informatica, Univ. degli Studi di Milano-Bicocca.

#### **Culture della materia abilitato a tenere esami**

- Insegnamento “Informatica Generale 2”, A.A. 2009/10 – 2015/16.  
Corso di Laurea in Lingue e Letterature Straniere, Univ. degli Studi di Bergamo,

### **Seminari all'interno di insegnamenti ufficiali**

- Insegnamento "Bioinformatica", A.A. 2008/09, 2009/10.  
Corso di Laurea Magistrale in Informatica, Univ. degli Studi di Milano-Bicocca.

### **Altre esperienze didattiche di livello universitario**

- Ciclo di seminari dal titolo "Elementi di Programmazione in C" all'interno del corso di formazione "Formazione di ricercatori e tecnici esperti nello sviluppo di metodologie per l'identificazione e il controllo delle infezioni degli animali - EPISUD" finanziato con il contributo del Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) nell'ambito del Programma Operativo Nazionale Ricerca e Competitività 2007-2013, Parco Tecnologico Padano S.r.l., 25-29 marzo 2013.
- Co-docenza nell'ambito dell'iniziativa "Learning Week", Univ. degli Studi di Milano-Bicocca, 8-12 febbraio 2010.

## Descrizione Estesa delle Attività di Ricerca

**Indicizzazione di collezioni di testi e applicazioni in bioinformatica.** La costruzione di strutture dati per l'indicizzazione di testi singoli o di collezioni di testi è un ambito classico di ricerca in informatica con importanti applicazioni che spaziano dalla ricerca rapida di sottostringhe, al confronto di testi, alla compressione. Uno degli indici di centrale importanza è l'*FM-index*<sup>1</sup>, una struttura dati compatta e efficiente per la ricerca di sottostringhe in un testo basata sulla *trasformata di Burrows–Wheeler (BWT)*<sup>2</sup> dello stesso. La BWT di un testo (o la sua generalizzazione a collezioni di testi), unita a informazioni aggiuntive come l'array Longest Common Prefix (LCP), ha diverse applicazioni, tra cui l'individuazione (efficiente) di fattori ripetuti all'interno del testo indicizzato (o di fattori comuni nella collezione di testi indicizzati). Per questo motivo, algoritmi efficienti per il calcolo di BWT e array LCP sono di primaria importanza.

In questa direzione ho contribuito alla progettazione e implementazione di un algoritmo per il calcolo di BWT e array LCP di grandi collezioni di stringhe che, basato su un approccio *divide-et-impera*, consente la parallelizzazione efficiente e che si è dimostrato competitivo con approcci allo stato dell'arte [J23, C14]. Inoltre, ho contribuito alla progettazione e implementazione di una nuova strategia di visita in memoria esterna per la costruzione contemporanea di BWT e array LCP di (grandi) collezioni di stringhe di differente lunghezza [J27]. La complessità computazionale nel caso peggiore dipende dal massimo valore dell'array LCP e, quindi, questo algoritmo è particolarmente competitivo nel caso di collezioni di stringhe dissimili tra loro, come mostrato nel confronto sperimentale di questo algoritmo con altri approcci allo stato dell'arte.

In bioinformatica, l'individuazione di sovrapposizioni significative tra stringhe (quindi fattori comuni a più testi) è un problema algoritmico centrale nell'assemblaggio di sequenze genomiche (una rassegna dei risultati più significativi è stata presentata in [J25]). Infatti, un approccio comunemente impiegato per l'assemblaggio si basa sulla visita di uno *string graph* – un grafo diretto che rappresenta in modo non ridondante tutte le sovrapposizioni significative tra i frammenti sequenziati. In questo ambito ho contribuito alla definizione, implementazione e sperimentazione di un algoritmo in memoria esterna – chiamato LSG – per la costruzione dello string graph di un (grande) insieme di frammenti [J20, J15, C11]. LSG evita il confronto pairwise esplicito tra i frammenti visitando unicamente la BWT della collezione dei frammenti e memorizza efficientemente in memoria esterna i risultati intermedi della visita, mentre mantiene in memoria centrale unicamente lo string graph finale, che è molto più piccolo dei risultati intermedi e dei dati in input. La sperimentazione su una collezione di 875 milioni di frammenti ha mostrato che LSG è stato in grado di costruire lo string graph utilizzando 50 volte meno memoria centrale dell'approccio correntemente allo stato dell'arte e richiedendo solo poco più di 2 volte il tempo impiegato da esso.

Parallelamente, ho contribuito alla definizione di un diverso algoritmo di visita della BWT di una collezione di frammenti che, unito a una ingegnerizzazione delle rappresentazioni di insiemi di frammenti che condividono sottostringhe, consente di ottenere il trade-off opposto [J21, C13]. Infatti, si è mostrato come questo nuovo algoritmo di costruzione di string graph riesce ad essere da 2.3 a 4.8 volte più veloce dell'approccio correntemente allo stato dell'arte, utilizzando al più 2.2 volte la memoria utilizzata da esso.

Con il termine *pangenomica* si intende l'analisi di sequenze basate su un insieme di genomi di riferimento correlati tra loro (ad es., una popolazione di individui). Questo consente di ottenere analisi più accurate, ma si scontra con l'assenza di modelli e algoritmi di uso consolidato per effettuare queste analisi. Una review dei modelli e degli algoritmi principali è stata presentata in [J28].

**Algoritmi per l'assemblaggio e l'analisi di sequenze genomiche e trascrittomiche.** L'assemblaggio di sequenze trascrittomiche è un problema che richiede di ricostruire le (numerosi) sequenze originali di

<sup>1</sup>Ferragina, P. "Opportunistic Data Structures with Applications." In: FOCS. IEEE CS, 2000.

<sup>2</sup>Burrows, M. A *block-sorting lossless data compression algorithm*. Tech. rep. DEC, 1994.

mRNA (chiamate *isoforme*), espresse dai geni di un organismo, a partire da loro frammenti. Le isoforme condividono numerose regioni comuni, in quanto i geni dell'organismo vengono sottoposti al meccanismo biologico di *splicing alternativo*, che espande l'insieme delle proteine codificate dal patrimonio genetico di un individuo. La caratterizzazione di tale meccanismo è un obiettivo di primaria importanza sia per la comprensione della fisiologia di un organismo sia per lo studio di numerose patologie in cui eventi aberranti di *splicing alternativo* sono un elemento determinante<sup>3</sup>. Il costo e il tempo necessario per la determinazione *in vitro* delle possibili isoforme rendono i metodi computazionali di predizione di isoforme l'unica strada di fatto perseguibile per lo studio su larga scala dello *splicing alternativo*.

A supporto della ricerca in questo ambito ho contribuito alla realizzazione di un algoritmo efficiente per la predizione delle isoforme. In particolare, in [J1] è stato proposto e sperimentato un algoritmo combinatorio per la predizione delle isoforme a partire dalle informazioni di allineamento di frammenti di sequenze espresse (chiamati *EST*). In presenza di *splicing alternativo*, i metodi tradizionali per l'allineamento di *EST* rispetto a una sequenza genomica di riferimento presentano importanti limiti. Di conseguenza ho disegnato e sviluppato un algoritmo [C4] in grado di sfruttare l'elevata ridondanza dei dati al fine di determinare una struttura di consenso della sequenza genomica, migliorando così l'affidabilità degli allineamenti delle *EST* su cui si basa la predizione di isoforme. Inoltre, ho ideato e realizzato un nuovo algoritmo efficiente di allineamento di sequenze espresse [J7]. Questo nuovo algoritmo di allineamento si basa sull'impiego efficiente di una struttura dati di indicizzazione chiamata *suffix-tree*<sup>4</sup> e sulla rappresentazione implicita degli allineamenti biologicamente plausibili tramite un grafo. L'integrazione delle implementazioni di questi tre algoritmi [J7, C7, B1] ha fornito alla ricerca biologica un tool completo, open source, accurato ed efficiente per la predizione di isoforme.

Recentemente, nuove tecnologie di sequenziamento – note con il nome di tecnologie di *Next Generation Sequencing (NGS)* – rendono disponibili grandi quantità di sequenze – chiamate *short-reads* – e consentono di studiare nuove specie di cui la sequenza genomica di riferimento non è nota o è di scarsa qualità. In questi casi, quindi, non si può ricorrere all'allineamento delle sequenze espresse rispetto alla sequenza genomica di riferimento. In questa direzione di ricerca, in [J11], si è presentato un duplice contributo. Da un lato si è definita formalmente una struttura a grafo, chiamata *grafo di splicing* e che rappresenta concisamente gli eventi di *splicing alternativo* desumibili dai frammenti ottenuti dalle tecnologie NGS, e si sono dimostrate un insieme di condizioni necessarie per la ricostruzione “corretta” del grafo di *splicing*. Queste condizioni rappresentano un limite teorico alla correttezza formale di un qualsiasi metodo di ricostruzione che opera unicamente sui frammenti osservati (le *short-read*). Dall'altro lato, si è sviluppato un metodo efficiente di ricostruzione del grafo di *splicing* a partire da grandi insiemi di *short-read*. L'efficienza del metodo è garantita attraverso l'utilizzo di *fingerprint* per indicizzare le *short-read* in tabelle hash e attraverso un'opportuna strategia di visita che consente di evidenziare gli eventi di *splicing alternativo* potenzialmente occorsi senza ricorrere a confronti pairwise tra le *short-read* (che richiederebbe una complessità computazionale che, seppur polinomiale, non sarebbe accettabile per l'analisi delle quantità di sequenze oggi disponibili). Si è dimostrata formalmente l'esistenza di alcune condizioni sufficienti per la ricostruzione corretta del grafo di *splicing* e si è verificata su dati simulati l'accuratezza del metodo anche in assenza delle condizioni formali di correttezza, dimostrandone così empiricamente la sua robustezza. Il confronto di questo metodo con un tool allo stato dell'arte ha evidenziato un'accuratezza comparabile tra le due predizioni. Tuttavia, il nostro metodo è in grado di fornire direttamente la descrizione degli eventi di *splicing alternativo* mentre il tool con cui ci si è confrontati, per far ciò, richiede l'utilizzo della sequenza genomica di riferimento, che non è sempre disponibile.

L'analisi di dati di espressione genica e proteica, ancorché ottenuti con tecniche analitiche tradizionali quali i microarray, pone tuttora interessanti problematiche di integrazione e di data mining. Per [J18, J16],

<sup>3</sup>Modrek, B. “A genomic view of alternative splicing.” *Nature Genetics* 30.1 (2002).

<sup>4</sup>Weiner, P. “Linear pattern matching algorithms.” In: *Switching and Automata Theory*. 1973.

ho messo a punto una metodologia di analisi, basata sul metodo presentato in<sup>5</sup>, per dati di espressione genica e dati di espressione proteica al fine di individuare pathway (o, più generalmente, insiemi di geni) che esibiscono differenze significative di espressione in due condizioni. I risultati ottenuti con questa analisi integrata su un dataset di linee cellulari normali e trasformate hanno portato ad individuare dei pathway (poi confermati sperimentalmente) non individuati sui singoli dati di espressione.

Il monitoraggio di terapie geniche richiede la determinazione dei siti di inserzione del DNA nel genoma dell'individuo in terapia. Tuttavia, i metodi esistenti mostrano limiti di precisione nell'individuazione delle inserzioni in regioni genomiche ripetitive. A supporto della ricerca in quest'ambito ho collaborato alla definizione di un algoritmo basato su grafi che migliora la precisione di metodi allo stato dell'arte, come mostrato da una rianalisi di studi esistenti [J24].

Gli esperimenti di sequenziamento RNA-Seq sono generalmente *genome-wide*, quindi producono dati riguardanti tutti i geni espressi in un individuo. Tuttavia, in alcuni casi, le analisi dei dati riguardano solo un insieme ristretto e predeterminato di geni (ad es., oncogeni) ma estrarre efficientemente le read relative a quei geni non è straightforward. Per questo motivo ho contribuito a sviluppare *Shark* [J26]: un tool basato su una nuova struttura di indicizzazione succinta per selezionare efficientemente tali read. Abbiamo mostrato che l'utilizzo di tale tool è in grado di velocizzare le analisi downstream senza variarne in modo significativo i risultati.

Il problema di rappresentare sequenze biologiche all'interno di modelli di machine learning è tuttora di interesse, in quanto è necessario darne una rappresentazione numerica che catturi il contenuto della sequenza. In questo ambito, ho contribuito [J30, C17, C16, C15] all'analisi di un metodo di fingerprinting di sequenze biologiche per catturare informazioni di overlap fra sequenze utilizzando alcune proprietà di una fattorizzazione di sequenze chiamata *Lyndon factorization*.

**Algoritmi di inferenza/assemblaggio di aplotipi e predizione di varianti geniche.** I problemi computazionali di inferenza/assemblaggio di aplotipi mirano a distinguere il patrimonio genetico ereditato da ciascun genitore (i due aplotipi, appunto) a partire da una rappresentazione parziale degli stessi. La determinazione degli aplotipi degli individui di una popolazione, ancora sostanzialmente proibitiva *in vitro*, è di primaria importanza per l'individuazione di possibili geni coinvolti nell'insorgenza di malattie o nella determinazione di caratteristiche dell'individuo<sup>6</sup>. I problemi computazionali di inferenza/assemblaggio di aplotipi si classificano in due categorie a seconda della tipologia dei dati sui quali operano<sup>7</sup>: i problemi di *haplotype assembly*, che operano su frammenti di ciascun aplotipo della coppia di aplotipi di un unico individuo (tipicamente derivanti da sequenziamenti next-generation), e i problemi di *phasing*, che operano su una rappresentazione congiunta delle coppie di aplotipi (chiamata genotipo) degli individui di una popolazione. Nel prosieguo verranno espone le mie attività di ricerca prima nel settore dell'assemblaggio di aplotipi da frammenti e, poi, le attività di ricerca nell'ambito di inferenza di aplotipi da genotipi. Infine verranno presentate le attività di ricerca su un problema che combina il processo di inferenza degli aplotipi con la determinazione di varianti geniche.

Nell'ambito dell'assemblaggio di aplotipi a partire da frammenti, il problema computazionale consiste nel ricostruire la coppia di aplotipi completi gestendo opportunamente gli errori di sequenziamento. Dal punto di vista combinatorio, questo problema è stato formulato come problema di ottimizzazione e diverse funzioni obiettivo sono state proposte<sup>8</sup>. Tra queste, la formulazione che ha mostrato di riuscire a ottenere risultati consistentemente più accurati è quella chiamata *Minimum Error Correction* (MEC). Il

<sup>5</sup>Väremo, L. "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods." *Nucleic Acids Research* 41.8 (2013).

<sup>6</sup>Altshuler, D. "Genetic mapping in human disease." *Science* 322.5903 (2008).

<sup>7</sup>Browning, S. "Haplotype phasing: existing methods and new developments." *Nature Reviews Genetics* 12.10 (2011).

<sup>8</sup>Geraci, F. "A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem." *Bioinformatics* 26.18 (2010).

problema MEC è APX-hard su istanze generiche e rimane NP-hard anche su istanze che non contengono gap<sup>9</sup>. L'investigazione teorica del problema ha comunque fornito approcci pratici al problema, specialmente mediante il disegno di algoritmi parametrici. In [C12] ho contribuito all'investigazione dell'intrattabilità computazionale del problema e di sue restrizioni, sia dal punto di vista parametrico, sia dal punto di vista della sua approssimabilità. In particolare, si è dimostrato che, assumendo la Unique Games Conjecture, MEC non è approssimabile a fattore costante (non appartiene a APX), che è approssimabile a fattore logaritmico nella dimensione dell'istanza, che MEC su istanze senza gap è fixed-parameter tractable quando parametrizzato nel numero di correzioni (cioè nell'ottimo del problema), e che MEC su matrici binarie è 2-approssimabile. Questi risultati sono stati inoltre estesi in [J17] e si è inoltre fornito una prima caratterizzazione degli aspetti di complessità computazionale di MEC al caso di genomi *poliploidi* (caso comune in molte piante, anche di forte interesse agricolo e commerciale come il frumento, la patata e il tabacco), cioè a genomi in cui ciascun cromosoma è presente in  $k$  copie omologhe (con  $k > 2$  fissato). La dimostrazione di alcuni di questi risultati si basa su una nuova caratterizzazione delle soluzioni ammissibili. Questa caratterizzazione è stata impiegata in [J14] per derivare un algoritmo esatto di una variante vincolata di MEC motivata dalle caratteristiche dei dati prodotti da tecnologie di sequenziamento di recente introduzione (tecnologie “third-generation”).

Nell'ambito dell'inferenza di aplotipi da genotipi, il processo di inferenza è guidato da un modello (*modello genetico*) di evoluzione e diffusione del patrimonio genetico degli individui della popolazione. A seconda delle caratteristiche della popolazione e dei dati in esame, sono stati proposti in letteratura diversi modelli genetici i quali, a loro volta, determinano diversi problemi computazionali<sup>10</sup>.

In questo ambito, nella mia attività di ricerca ho studiato il problema *Pure Parsimony Xor Haplotyping*, per il quale ho disegnato algoritmi risolutivi esatti (in tempo polinomiale per istanze ristrette e parametrici per istanze generali), di approssimazione e euristici [J3, C3]. Il problema, motivato biologicamente<sup>11,12</sup>, ha una struttura combinatoria che mostra profonde relazioni con problemi su grafi (come *Graph Realization*<sup>13</sup>) di interesse più generale.

La rappresentazione della storia evolutiva di caratteri genetici e delle relazioni di parentela tra gli individui di una popolazione fornisce informazioni aggiuntive per l'inferenza di aplotipi. Una rassegna di queste rappresentazioni è stata prima presentata in [J2] e, in versione rivista e aggiornata, in [B2]. In particolare, in letteratura si è mostrato che l'inferenza di aplotipi in popolazioni di cui si conoscono i rapporti di parentela che intercorrono fra gli individui (*pedigree*) è precisa e affidabile<sup>14</sup>.

In [J5, C6] è stato ideato e sperimentato un metodo per l'inferenza di aplotipi in pedigree e in presenza delle due tipologie di eventi di variazione genetica più comuni nella trasmissione del patrimonio genetico da genitore a figlio. Questo algoritmo è basato su una riduzione polinomiale del problema di inferenza di aplotipi in un problema di teoria dei codici, chiamato *Nearest Codeword Problem*<sup>15</sup>, in modo tale che eventi di variazione corrispondano a violazioni di vincoli di parità di un codice lineare.

In [J8, C8] si è proposto un'estensione della formulazione combinatoria del problema di inferenza di aplotipi in pedigree che include la possibile presenza di errori nei genotipi osservati sperimentalmente. Gli errori di genotipizzazione possono pregiudicare i risultati delle analisi successive ma non sono considerati

<sup>9</sup>Cilibiasi, R. “The complexity of the single individual SNP haplotyping problem.” *Algorithmica* 49.1 (2007).

<sup>10</sup>Gusfield, D. “Haplotype inference.” In: *CRC Handbook on Bioinf.* 2006.

<sup>11</sup>Barzuza, T. “Typing without calling the allele: A strategy for inferring SNP haplotypes.” *European Journal of Human Genetics* 13.8 (2005).

<sup>12</sup>Climer, S. “How frugal is mother nature with haplotypes?” *Bioinformatics* 25.1 (2009).

<sup>13</sup>Tutte, W. “An algorithm for determining whether a given binary matroid is graphic.” *Proceedings of the American Mathematical Society* 11.6 (1960).

<sup>14</sup>Zhang, K. “A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers from general pedigrees.” *Genetic Epidemiology* 30.5 (2006).

<sup>15</sup>Ausiello, G. *Complexity and approximation: Combinatorial optimization problems and their approximability properties*. Springer-Verlag, 1999.

da molti metodi esistenti di inferenza di aplotipi perché difficili da individuare e da gestire. Per questa formulazione del problema (che si dimostra essere computazionalmente difficile, in particolare APX-hard) si è proposto un algoritmo risolutivo dimostratosi efficiente e accurato in una sperimentazione estensiva su dati reali e simulati e in un suo confronto con altri tool riconosciuti in letteratura. Questo algoritmo è basato sulla riduzione polinomiale dell'istanza del problema di inferenza di aplotipi a un'istanza del problema di *Soddisfacibilità Booleana* (SAT) combinata con un'efficiente ricerca binaria all'interno dello spazio delle soluzioni ammissibili. La riduzione è stata disegnata per produrre formule booleane con caratteristiche adatte alle strategie di esplorazione implementate nei risolutori SAT allo stato dell'arte.

Come estensione di questo algoritmo, ho proposto un metodo per la determinazione di *varianti geniche* degli individui di una popolazione, cioè le diverse “forme” con cui un gene può essere presente in ciascun individuo, a partire dai genotipi di tutti gli individui e dalle varianti geniche possedute da alcuni individui della popolazione. Il metodo, presentato in [C9], integra due ulteriori fasi all'algoritmo precedente. La prima fase mira a determinare un'insieme di associazioni fra aplotipi e varianti geniche su alcuni individui di cui si conoscono i genotipi SNP e le varianti geniche, mentre la seconda fase applica le associazioni per determinare le plausibili varianti geniche degli individui di cui sono noti solamente i genotipi SNP. In particolare, la prima fase calcola un insieme di associazioni fra aplotipi SNP e varianti geniche che minimizza gli errori di associazione (cioè individui per cui le varianti geniche osservate non corrispondono a quelle associate ai loro aplotipi), tramite una formulazione di programmazione lineare intera (ILP). La seconda fase, invece, determina le possibili varianti geniche di ciascun individuo di cui è noto il solo genotipo SNP mediante un meccanismo ispirato al *majority voting*. Questo metodo è stato valutato sui dati reali di una popolazione bovina di 1655 animali ottenuti grazie al supporto dell'Associazione Nazionale Allevatori Razza Bruna (ANARB). L'accuratezza della predizione è stata valutata mediante una strategia di validazione incrociata e il tasso di errore mediano si è rivelato migliore del 21% di quello ottenuto tramite un altro metodo allo stato dell'arte. Confronti fra i metodi allo stato dell'arte per la predizione di varianti geniche sono stati presentati in [J19].

La pandemia provocata dal virus SARS-CoV-2 ha dimostrato sul campo che i dati di sequenziamento possono contribuire efficacemente in diversi ambiti nella gestione e nella risposta a questi eventi. Un task comune alle diverse applicazioni del sequenziamento di SARS-CoV-2 è l'identificazione del genotipo di un sample (quindi l'insieme delle sue variazioni rispetto al riferimento) e, in seguito, la sua classificazione in una delle possibili varianti. Per assistere i ricercatori impegnati in questo compito che non abbiano una specifica formazione in bioinformatica, in [J29] ho contribuito all'ideazione e realizzazione di una pipeline di analisi integrata in un'applicazione di facile utilizzo e alla realizzazione di un metodo per la creazione di un catalogo di variazioni genetiche rappresentativo della variabilità genetica dei vari sample virali depositati nelle banche dati pubbliche.

**Ricostruzione vincolata e confronto vincolato di sequenze.** Il confronto di sequenze è un tema di ricerca con importanti applicazioni in bioinformatica e altre discipline, come la compressione di dati e la gestione di testi<sup>16</sup>. In quest'area di ricerca, ho contribuito [J4] all'analisi di un problema di determinazione della più lunga sottosequenza comune a due stringhe soggetta a vincoli di occorrenza dei simboli e di inclusione di sottosequenze (*DC-LCS problem*). Tale problema è la naturale formulazione di problematiche relative al confronto di genomi di specie differenti, al fine di determinare, ad esempio, misure di distanza evolutiva fra le specie in esame<sup>17</sup>. Il contributo di questo lavoro è duplice. Innanzitutto, fornisce un algoritmo esatto di tipo parametrico per il problema nel caso in cui il parametro fissato è la lunghezza della soluzione ottima. In secondo luogo, invece, dimostra la difficoltà parametrica ( $W[1]$ -hardness) di un'importante restrizione del problema DC-LCS, implicando lo stesso risultato per il problema generale.

<sup>16</sup>Gusfield, D. *Algorithms on strings, trees and sequences*. Cambridge University Press, 1997.

<sup>17</sup>Sankoff, D. “Genome rearrangement with gene families.” *Bioinformatics* 15.11 (1999).



Il problema di ricostruzione di sequenze a partire da loro frammenti/short-read è stato spesso modellato in letteratura come un problema di copertura dei vertici di un grafo diretto e aciclico (DAG) mediante percorsi. In questa modellizzazione, ciascun vertice rappresenta un frammento e un arco collega due frammenti se essi sono significativamente sovrapposti. Di conseguenza, un percorso massimale all'interno del grafo rappresenta una sequenza nucleotidica che potrebbe aver dato origine ai frammenti che "copre". Il successo di questa modellizzazione è anche dovuto alla possibilità di risolvere in tempo polinomiale il problema combinatorio di copertura di un DAG mediante percorsi. Alcune informazioni aggiuntive (quali, ad esempio, la presenza di frammenti *paired-end*) potrebbero indurre vincoli durante la ricostruzione dei percorsi e, quindi, migliorare l'accuratezza della predizione, ma sono spesso ignorate o sfruttate solo in una seconda fase per eliminare ricostruzioni non compatibili con i vincoli osservati. Di conseguenza, la formulazione e lo studio della complessità computazionale di problemi di copertura vincolata di DAG mediante percorsi assume un ruolo di primo piano per il miglioramento dei metodi di ricostruzione di sequenze.

In quest'ambito, ho contribuito [J13, C10] allo studio della complessità computazionale di problemi di copertura di un DAG mediante percorsi vincolati dalla presenza di *required pair* (che modellano i frammenti *paired-ends*). In particolare, si sono formulati due problemi di copertura e se ne è studiata la loro complessità computazionale. Il primo – chiamato *Minimum Path Cover with Required Pairs (MinPCRP)* – richiede di calcolare un insieme di minima cardinalità di percorsi tale che copre tutti i vertici del DAG in input e che per ogni *required pair* in input esista un percorso che lo copra. Di questo problema si è presentato un confine netto per la sua complessità computazionale dimostrandone la trattabilità quando la soluzione è composta da 2 percorsi e l'intrattabilità (NP-hardness) già quando la soluzione è composta da soli 3 percorsi. Il secondo problema che si è formulato – chiamato *Maximum Required Pairs with a Single Path* – richiede di calcolare un percorso del DAG in input che soddisfa il massimo numero di *required pair* in input. Questa formulazione corrisponde a una strategia greedy per il problema MinPCRP in cui ad ogni iterazione si copre il maggior numero di *required pair* non ancora coperti. Tuttavia, si è dimostrata l'intrattabilità del problema (NP-hardness) che si estende anche in senso parametrico (W[1]-hardness) quando il parametro è il numero di *required pair* coperti dal percorso. Dal lato positivo, per questo problema si è disegnato un algoritmo parametrico quando il parametro è il numero di *required pair* "sovrapposti", parametro che potrebbe avere un valore limitato in alcune delle applicazioni alla ricostruzione di sequenze a partire da *paired-end*.

**Complessità e algoritmi parametrici.** Durante la mia attività di ricerca, ho analizzato ulteriori problemi combinatori concentrandomi sullo studio della loro complessità computazionale e la loro risoluzione efficiente, con particolare riguardo agli aspetti di *complessità parametrica* e, ove possibile, al disegno di *algoritmi parametrici*. Informalmente, lo studio della complessità parametrica (*parameterized complexity*) è un approccio, proposto da Downey and Fellows<sup>18</sup>, per analizzare e catalogare con maggior precisione la complessità computazionale di problemi NP-hard individuando un aspetto dell'istanza, il *parametro*, da cui dipende l'intrattabilità del problema stesso. Il disegno di algoritmi parametrici mira a isolare l'esplosione esponenziale dei tempi di calcolo al solo parametro del problema e fornisce, così, algoritmi per quanto possibile efficienti in pratica qualora il valore del parametro sia di modesta entità rispetto alla dimensione dell'istanza.

Un problema combinatorio di recente interesse è il problema *k-anonymity*, relativo alla pubblicazione di dati sensibili preservando la privacy individuale<sup>19</sup>. In particolare, questo problema richiede la cancellazione del minimo quantitativo di dati sensibili all'interno di un dataset da rendere pubblico di modo che, in seguito alla cancellazione, i dati relativi a ciascun individuo siano indistinguibili da quelli relativi ad, almeno,  $k - 1$  altri individui. In [J10, C5] ho contribuito allo studio sistematico della complessità compu-

<sup>18</sup>Downey, R. *Parameterized Complexity*. Springer-Verlag, 1999.

<sup>19</sup>Samarati, P. "Generalizing data to provide anonymity when disclosing information." In: *Principles of Database Systems*. 1998.

tazionale di questo problema, dimostrando come lo stesso sia intrattabile anche nel senso parametrico in diversi casi di rilevanza pratica e rimasti aperti in lavori precedentemente apparsi in letteratura.

Il problema di integrare numerosi schemi concettuali in un unico schema che rappresenti l'informazione presente in ognuno è un problema rilevante per assicurare un'alta qualità nei sistemi informativi aziendali. In [J12] ho contribuito alla definizione del problema di assistere a questo processo di integrazione come problema di *clusterizzazione vincolata* e per questo problema ho definito un algoritmo euristico basato su una formulazione ILP. La sperimentazione su un insieme di schemi concettuali derivanti dal sistema informativo fiscale italiano ha mostrato che la qualità degli schemi concettuali integrati a partire dalla clusterizzazione proposta dall'algoritmo è superiore a quella ottenuta da una clusterizzazione effettuata manualmente da un esperto di settore.

In letteratura è stato recentemente proposto<sup>20</sup> il problema di determinare il massimo insieme di cammini disgiunti tra una coppia di nodi in grafi colorati. Questo problema, che estende il classico problema di calcolare la connettività sui vertici di un grafo, è motivato da studi sulla struttura di *multi-relational social network* ove, tramite i vari colori, vengono considerate e rappresentate diverse tipologie di relazioni fra le entità della rete. In [J9] ho contribuito allo studio della complessità computazionale del problema dimostrando, da un lato, l'inapprossimabilità del problema con un fattore minore del numero di colori del grafo e la sua intrattabilità in senso parametrico ( $W[1]$ -hardness), mentre, dall'altro lato, l'esistenza di un algoritmo parametrico per il problema dove il parametro è rappresentato dalla lunghezza e dal numero di cammini disgiunti.

**Misure di difficoltà in programmazione genetica.** Parte della mia attività di ricerca si è concentrata sullo studio della difficoltà di problemi di ottimizzazione affrontati con una tecnica metaeuristica nota come *programmazione genetica*<sup>21</sup>. La programmazione genetica è una tecnica che consente di generare automaticamente (possibilmente) buone soluzioni per problemi di ottimizzazione di cui non si è in grado di disegnare un algoritmo risolutivo diretto. Teoricamente la programmazione genetica è in grado di calcolare una soluzione ottima in un numero sufficientemente grande di iterazioni<sup>22</sup>. In pratica, predire l'efficacia di questa metaeuristica su un determinato problema (cioè la capacità di generare una "buona" soluzione al problema con uno sforzo computazionale "contenuto") è un compito complesso.

In quest'ambito, in [J6, C2, C1] ho proposto, analizzato e confrontato alcune misure quantitative della difficoltà di problemi per la programmazione genetica sulla base delle caratteristiche di una rappresentazione dello spazio delle soluzioni di un problema di ottimizzazione.

Ultimo aggiornamento: 21 ottobre 2022

---

<sup>20</sup>Wu, B. "On the maximum disjoint paths problem on edge-colored graphs." *Discrete Optimization* 9.1 (2012).

<sup>21</sup>Koza, J. *Genetic programming: On the programming of computers by means of natural selection*. MIT Press, 1992.

<sup>22</sup>Langdon, W. *Foundations of genetic programming*. Springer, 2002.