Informazioni Personali

COGNOME	Frasca
NOME	Marco
DATA DI NASCITA	18.09.1979
E-MAIL	marco.frasca at unimi dot it
WEB-SITE	http://frasca.di.unimi.it

Marco Frasca
Curriculum vitae accademico aggiornato al 07.2019

Indice

1 Breve Biografia

1.1 Posizione attuale

Ricercatore tipo A
Dipartimento di Informatica, Università degli Studi di Milano

1.2 Stato di servizio

- Maggio 2017 - Oggi. Ricercatore di tipo A presso il dipartimento di Informatica dell'Università degli Studi di Milano. Durata triennale.
- Aprile 2013 - Marzo 2017. Assegno di ricerca di tipo A presso il dipartimento di Informatica dell'Universià degli Studi di Milano, dal titolo "Algoritmi neurali basati su grafi per l'analisi di reti biologiche"
- Giugno 2012 - Marzo 2013. Assegno di ricerca di tipo B presso il dipartimento di Bioscienze dell'Università degli Studi di Milano, dal titolo "Sviluppo di algoritmi e software per l'analisi dello stato della cromatina in genomi di eucarioti"
- Gennaio 2009 - Dicembre 2011. Borsa di studio triennale per il finanziamento di dottorati di ricerca presso il Dipartimento di Informatica dell'Università degli Studi di Milano
- Luglio 2008 - Ottobre 2008. Contratto di inserimento FIXO presso l'azienda Metoda S.P.A., Salerno, per lo studio e l'implementazione di protocolli di sicurezza, certificazione e mutua autenticazione per il progetto SAFE: Sistema di Anamnesi in Fase di Emergenza.
- Gennaio 2007- Dicembre 2007. Borsa di studio annuale per il finanziamento della formazione di docenti di primo e secondo grado presso l'Università degli Studi di Salerno.
- Gennaio 2006 - Settembre 2006. Contratto di lavoro a tempo indeterminato presso l'azienda Texa S.P.A., Parma, per lo sviluppo di sistemi in linguaggio C per l'autodiagnosi.

1.3 Formazione

- Gennaio 2009 - Dicembre 2011. Studente del corso di dottorato in informatica XXIV ciclo, con borsa di studio triennale, presso il Dipartimento di Informatica dell'Università degli Studi di Milano. Titolo conseguito il 6 marzo 2012. Tesi discussa: "Graph-based approaches for imbalanced data in functional genomics".
  Relatore: Prof. Alberto Bertoni
  Correlatore: Prof. Giorgio Valentini
- Ottobre 2006 - Maggio 2008. Corso di abilitazione, con borsa di studio annuale, all'insegnamento della matematica nelle scuole, classe A047, presso la scuola inter-universitaria di Salerno (SICSI). Abilitazione conseguita il 22 maggio 2008 con votazione finale 80/80.
- Maggio 2005. Laurea quinquennale (vecchio ordinamento) in informatica presso l'Università degli Studi di Salerno. Tesi discussa: "L'operazione di composizione nella famiglia dei codici massimali prefissi".
  Relatore: Prof.ssa Clelia De Felice.
  Votazione finale 110/110 con lode.

1.3.1 Soggiorni presso centri di ricerca esteri e partecipazione a scuole internazionali

Soggiorni presso centri di ricerca esteri

- 1 maggio – 4 agosto 2019. Soggiorno di ricerca presso il Department of Computer Science della Royal Holloway University of London.
- 1 settembre – 30 settembre 2017. Soggiorno di ricerca presso il Dipartimento di Medicina del Tanz Centre for Research in Neurodegenerative Diseases, Università di Toronto, per una collaborazione con il Prof. Ming Zhang circa l'applicazione di un nuovo metodo computazionale per l'individuazione dei geni associati al morbo di Alzheimer.
- 4 settembre – 1 ottobre 2016. Soggiorno di ricerca presso l'Istituto di Biologia Molecolare (IMB) della Johannes Gutenberg University of Mainz, per una collaborazione con il Computational Biology and Data Mining Group diretto dal Prof. Miguel Andrade Navarro.
- 22 settembre 2014 - 6 ottobre 2014. Visiting Researcher presso il Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada. In collaborazione con il Prof. Quaid Morris, direttore del laboratorio, si sono studiati metodi di correzione pre-predizione delle etichette per algoritmi di label propagation.

Partecipazione a scuole di dottorato di alta formazione

- 14-19 agosto 2011. "From Data to Models in Biological Systems", Kandersteg (Switzerland). Organizzata dallo Swiss Institute of Bioinformatics (SIB).
- 12-19 giugno 2010. "Statistical and Machine Learning Methods in Computational Biology", Lipari. Organizzata dalla Jacob T. Schwartz International School fo Scientific Research.

1.3.2 Collaborazioni con gruppi di ricerca nazionali e internazionali

Collabora con diversi gruppi di ricerca ed enti nazionali e internazionali, tra cui la Royal Holloway University di Londra, la University Charité di Berlino, la Johannes Gutenberg University of Mainz, la Fondazione Centro San Raffaele e l'Istituto Nazionale dei Tumori. Di seguito un elenco dettagliato.

Collaborazioni con gruppi di ricerca internazionali:

- *In corso.* PaccanaroLab del Department of Computer Science della Royal Holloway University of London, circa lo studio di metodi di computazionali per l'individuazione dei geni responsabili di patologie genetiche complesse, con particoloare attenzione al trasferimento di nuova informazione da organismi modello vicini filogeneticamente all'uomo. La collaborazione è finalizzata alla sottomissione di un progetto di ricerca su bando competitivo. Inoltre, oggetto della collaborazione è anche lo sviluppo di metodi di apprendimento automatico per la predizione di nuovi potenziali effetti collaterali di farmaci già sul mercato. La collaborazione ha già portato alle pubblicazioni [17, 34].
- *In corso.* Partecipazione come membro del laboratorio AnacletoLab di UNIMI alla challenge internazionale CAFA3 (Critical Assessment of Functional Annotation) nell'ambito dello Special Interest Group "Protein Function Prediction" di ISCB (International Society of Computational Biology). Lo Special Interest Group riunisce i principali gruppi di ricerca internazionali per la predizione della funzione delle proteine con metodi computazionali[1], tra

[1] `http://biofunctionprediction.org/`

cui il Paccanaro Lab[2], il Rost Lab[3] e il Gough Lab[4]. I risultati della collaborazione sono stati sottomessi a rivista internazionale di riferimento del settore [1].

- *In corso.* Collaborazione con il Computational Biology and Data Mining Group della Johannes Gutenberg University of Mainz, diretto dal Prof. Miguel Andrade Navarro, circa lo sviluppo e l'applicazione di algoritmi per l'individuazione delle sorgenti di dati più informative e la loro integrazione per specifiche patologie genetiche, al fine di determinarne i geni responsabili. La collaborazione ha prodotto sinora la pubblicazione [25].

- *In corso.* Collabora con l'Uninversità Ca'Foscari di Venezia, gruppo del professor Marcello Pelillo, per lo sviluppo di algoritmi di teoria dei giochi per la classificazione multitask applicata alla predizione della funzione proteica. I risultati della ricerca sono stati pubblicati in [8].

- *In corso.* Colaborazione con l'ECLT (European Center for Living Technologies) per lo sviluppo di metodi basati sulla teoria dei giochi per l'analisi di reti biomolecolari con applicazioni nell'ambito della biologia molecolare di base e della "Network medicine".

- *Novembre 2014 - Febbraio 2015.* In collaborazione con il gruppo del Prof. Peter N. Robinson, direttore del Computational Biology Group presso il Max Planck Insititute for Molecular Genetics, University Charité di Berlino, studiati metodi di correzione gerarchica post-predizione per tassonomie gerarchiche. La collaborazione ha prodotto la pubblicazione [32].

Collaborazioni con aziende e istituti di ricerca nazionali:

- Gennaio 2016 - oggi. In collaborazione col Dip. di Neurologia della Fondazione Centro San Raffaele di Milano, analisi e sviluppo di metodi automatici per la predizione della risposta al farmaco FINGOLIMOD in pazienti affetti da sclerosi multipla. La collaborazione ha portato al deposito di un Brevetto presso l'Ufficio Italiano Brevetti di una signature di 21 SNP associati alla risposta al farmaco Fingolimod (tuttora in validazione).

- 2015 - 2016. In collaborazione con l'Istituto Nazionale dei Tumori di Milano, con il gruppo di ricerca del Dott. Tommaso Dragani, in corso lo studio di algoritmi di apprendimento automatico per la predizione del fenotipo tumorale al fegato e al polmone nell'organismo *Mus Musculus*.

1.3.3 Seminari sulla attività di ricerca

Ha tenuto i seguenti seminari

- "Parametric Hopfield networks for classification with unbalanced data. From single- to multitask learning". Ciclo di due seminari presso il Department of Computer Science della Royal Holloway University of London, 23/30 luglio 2019.

- "Machine Learning for Bioinformatics and Personalized Medicine: a survey of my research activity at the Computer Science Dept UNIMI", Dipartimento di Informatica, Università degli Studi di Milano, 20 marzo 2017.

- "Algoritmi neurali basati su grafi per l'analisi di reti biologiche", Dipartimento di Informatica, Università degli Studi di Milano, 09 aprile 2015.

[2] `http://www.paccanarolab.org/`
[3] `https://www.rostlab.org/`
[4] `http://home.cc.umanitoba.ca/~kmgough/`

1.3.4 Riconoscimenti e borse di studio

- Aprile 2013 - oggi. Vincitore di un assegno di ricerca di tipo A presso il dipartimento di Informatica dell'Universià degli Studi di Milano, dal titolo "Algoritmi neurali basati su grafi per l'analisi di reti biologiche".
- Giugno 2012 - Marzo 2013. Vincitore di assegno di ricerca di tipo B presso il dipartimento di Bioscienze dell Università degli Studi di Milano, dal titolo "Sviluppo di algoritmi e software per l'analisi dello stato della cromatina in genomi di eucarioti"
- Gennaio 2009 - Dicembre 2011. Vincitore di una borsa di studio triennale per il finanziamento di dottorati di ricerca presso il Dipartimento di Informatica dell'Università degli Studi di Milano
- Gennaio 2007- Dicembre 2007. Vincitore di borsa di studio annuale per il finanziamento della formazione di docenti di primo e secondo grado presso la S.I.C.S.I. dell'Università degli Studi di Salerno.

1.3.5 Partecipazione in qualità di relatore a congressi di interesse nazionale e internazionale

- 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Wurzburg, Germania, 16-20 settembre 2019. Presentazione del lavoro [21].
- 2019 9th 9th International Conference on Biomedical Engineering and Technology (ICBET 2019) March 28-30, 2019, Tokyo, Japan. Presentazione del lavoro [24].
- 2016 15th International Workshop on Data Mining in Bioinformatics (BIOKDD '16), all'interno della ACM SIGKDD 2016 Conference on Knowledge Discovery and Data Mining,13-17 August, San Francisco, USA. Presentazione del lavoro [29].
- 2016 International Work-Conference on Bionformatics and Biomedical Engineering (IWBBIO 2016), Granada, Spagna, 20-22 aprile 2016. Presentazione del lavoro [30].
- 2013 International Joint Conference on Neural Networks (IJCNN 2013), Dallas, Texas, 4-9 agosto 2013. Presentazione del lavoro [35].
- 2012 22th Italian Workshop on Neural Networks, Vietri sul Mare, Salerno, Italy, 17-19 maggio 2012. Presentazione del lavoro [36].
- 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Atene, Grecia, 5-9 settembre 2011. Presentazione del lavoro [37].
- 2011 Eighth Annual Meeting of the Bioinformatics Italian Society, Pisa, Italia, 20-22 giugno 2011. Presentazione del lavoro [38].
- 2010 20th Italian Workshop on Neural Networks, Vietri sul Mare, Salerno, Italia, 27-29 maggio 2010. Presentazione del lavoro [39].

1.3.6 Partecipazione a challenge internazionali

Ha partecipato come membro del gruppo di Bioinformatica e Biologia computazionale (AnacletoLab) del Dipartimento di Informatica dell'Università degli Studi di Milano alla terza edizione della challenge internazionale Critical Assessment of protein Function Annotation (CAFA), maggiore challenge internazionale per la predizione delle funzioni proteiche, che vede impegnati più di 100 gruppi di ricerca. Ha contribuito con l'applicazione di diversi dei modelli di apprendimento sviluppati [5, 10, 11, 18], risultando con i membri di AnacletoLab tra gli autori della susseguente pubblicazione [1].

2 Attività editoriale e Organizzazione di Conferenze/workshop internazionali:

Svolge/ha svolto numerose revisioni di lavori per riviste internazionali e conferenze nazionali e internazionali di machine learning, reti neurali e intelligenza artificiale. Di seguito un elenco delle principali.

Riviste internazionali:

- Scientific Reports
- Expert Systems with Application
- Neural Networks
- Neurocomputing
- IEEE Transactions on Neural Networks and Learning Systems
- BMC Bioinformatics
- Information Sciences

Conferenze internazionali

- International Joint Conference on Artificial Intelligence (IJCAI) 2015
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2014.
- International Conference on Artificial Neural Networks (ICANN), 2014.
- International Conference on Artificial Neural Networks (ICANN), 2013.

Svolge attività di revisione di progetti per il *Natural Sciences and Engineering Research Council of Canada* (NSERC), nell'area *Discovery Grant proposal*[5].

È revisore esterno di progetti di ricerca pluriennali dell'area Innovational Research Incentives Scheme (VIDI programme) presentati da giovani ricercatori (3–8 anni dal dottorato) presso la Netherlands Organisation for Scientific Research (NWO, the Dutch Research Council)[6].

È membro del Comitato di Programma dei seguenti convegni internazionali:

[5] http://www.nserc-crsng.gc.ca/index_eng.asp

[6] http://www.nwo.nl/en/funding/our-funding-instruments/nwo/innovational-research-incentives-scheme/vidi/index.html

- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2019), October 13-16, 2019 Pittsburgh, PA, USA[7]
- International Conference on Artificial Neural Networks (ICANN 2019) Munich, Germany,17–19 Semptember 2019[8]
- International Conference on Signal Processing and Machine Learning (SPML 2019), Hangzhou, China, November 27-29, 2019[9]
- 2019 5th International Conference on Biomedical and Bioinformatics Engineering (ICBBE 2019) Shanghai, China, 2019.
- 16th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2019), Bergamo, Italy, 4-6 September 2019[10]
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019) Brighton, UK, 12-17 May 2019[11]
- First International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2019), Barcelona, Spain, 20-22 March 2019[12]
- Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB 2018), Caparica, Portugal, September 6-8, 2018[13]
- International Conference on Signal Processing and Machine Learning (SPML 2018), Shanghai, China, November 28-30, 2018[14]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2018), Aalborg, Denmark, September 17-20, 2018[15]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2017), Tokyo, Japan, September 25-28, 2017[16]
- IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2016), Vietri sul Mare, 13–16 settembre 2016[17].

È Associate Editor dell'International Journal of Neural Networks (ISSN:2249-2763 print version,E-ISSN: 2249-2771 electronic version, DOI:10.9735/2249-2763).

È membro della Bioinformatics Italian Society (BITS) dal 2017.
È stato membro della International Neural Network Society (INNS) dal maggio 2013 al dicembre 2014.

Inoltre è stato session chair della conferenza IJCNN (International Joint Conference of Neural Network) 2013[18], nella sessione dedicata alla Bioinformatica.

[7] https://www.ieeemlsp.cc/
[8] https://e-nns.org/icann2019/
[9] http://www.spml.net/
[10] www.cibb2019.it
[11] https://2019.ieeeicassp.org/
[12] https://drive.google.com/file/d/1hLl_rXea028z6q9GJIxK5UK0LGrmJmR7/view
[13] https://eventos.fct.unl.pt/cibb2018/pages/organization
[14] http://www.spml.net
[15] http://mlsp2018.conwiz.dk/home.htm
[16] https://signalprocessingsociety.org/blog/mlsp-2017-2017-ieee-international-workshop-machine-learning-signal-processing
[17] http://www.conwiz.dk/cgi-all/mlsp2016/list_pc.pl
[18] http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=31079

3 Attività Didattica, di Tutorato e Servizio agli Studenti

3.1 Attività di supporto alla didattica

- A.A. 2018-2019. Affidamento del modulo di Laboratorio Basi di Dati (48 ore). Dipartimento di Informatica, Università degli Studi di Milano.
- 2018. Laboratorio Data Mining e Machine Learning (25 ore). Master di secondo livello in Data science for economics, business and finance, Università degli Studi di Milano.
- A.A. 2017-2018. Affidamento del modulo di Laboratorio Basi di Dati (48 ore). Dipartimento di Informatica, Università degli Studi di Milano.
- A.A. 2015-2016. Affidamento del modulo di Laboratorio di Algoritmi (44 ore). Dipartimento di Matematica, Università degli Studi di Milano.
- A.A. 2015-2016. Esercitazioni di Statistica e Analisi dei dati. (4 ore). Dipartimento di Informatica, Università degli Studi di Milano.
- A.A. 2015-2016. Laboratorio di Algoritmi e Strutture Dati (12 ore). Dipartimento di Informatica, Università degli Studi di Milano.
- A.A. 2014-2015. Affidamento del modulo di Laboratorio di Algoritmi (44 ore). Dipartimento di Matematica, Università degli Studi di Milano.
- A.A. 2013-2014. Affidamento del modulo di Laboratorio di Algoritmi (44 ore). Dipartimento di Matematica, Università degli Studi di Milano.
- A.A. 2013-2014. Informatica per Biotecnolgie (50 ore). Dipartimento di Bioscienze, Università degli Studi di Milano.
- A.A. 2012-2013. Affidamento del modulo di Laboratorio di Algoritmi (44 ore). Dipartimento di Matematica, Università degli Studi di Milano.
- A.A. 2011-2012. Affidamento del modulo di Laboratorio di Algoritmi (44 ore). Dipartimento di Matematica, Università degli Studi di Milano.
- A.A. 2011-2012. Ha svolto 16 ore di attività di tutoraggio per il corso di Informatica B, presso la Scuola di Ingegneria Industriale del Politecnico di Milano.
- A.A 2010-2011. Affidamento del modulo di Laboratorio di Algoritmi e Strutture Dati, turno serale (48 ore). Dipartimento di Informatica, Università degli Studi di Milano.

Per tutti gli affidamenti indicati in precedenza ha partecipato alle commissioni d'esame. Ha inoltre svolto attività di supporto per i seguenti esami:

- A.A. 2015-2016. Statistica e Analisi dei dati (30 0re). Dipartimento di Informatica, Università degli Studi di Milano.
- A.A. 2011-2012. Laboratorio di Algoritmi (20 ore). Dipartimento di Matematica, Università degli Studi di Milano.

3.2 Correlatore di tesi di laurea triennale e magistrale.

È stato correlatore di tesi di laurea triennale e magistrale presso il Dipartimento di Informatica dell'Università degli Studi di Milano:

- 8 tesi di laurea triennale (2 in corso)
- 3 tesi di laurea magistrale

Gli argomenti studiati nelle tesi hanno riguardato prevalentemente lo sviluppo di modelli di apprendimento basati su reti di Hopfield con parametri multipli e su restricted Boltzman machines, la compressione di reti neurali, l'analisi di metodi si selezione di esempi negativi in problemi di Positive-Unlabeled Learning e l'implementazione di servizi web per la computazione di misure di similarità semantica in tassonomie strutturate come grafi.

4 Partecipazione a Progetti di Ricerca

Ha presentato i seguenti progetti di ricerca su bandi competitivi:

- Progetto PRIN 2017.
  `Esito`. Vincitore.
  `Titolo`. Multicriteria Data Structures and Algorithms: from compressed to learned indexes, and beyond.
  `Ente finaziatore`. MIUR.
  `Ruolo svolto`. Responsabile dell'unità di ricerca milanese (Dipartimento di Informatica 'Giovanni degli Antoni').
  `Ateneo coordinatore`. Università di Pisa.
  `Durata`: 3 anni, a partire da 29/08/2019.

- Progetto SEED 2019.
  `Esito`. In valutazione.
  `Titolo`. Valutazione personalizzata del rischio trombotico ed emorragico dei pazienti con fibrillazione atriale.
  `Ente finaziatore`. UNIMI.
  `Ruolo svolto`. Responsabile (CUD) della U.D. Dipartimento di Informatica 'Giovanni degli Antoni'.
  `Durata`: 1 anno.

Gestisce/ha gestito i seguenti progetti di ricerca:

- `Titolo`. Hierarchical classification algorithms in biomedical taxonomies.
  `Ente finanziatore`: UNIMI - Piano di Sostegno alla Ricerca 2015-2017
  `Ruolo`: responsabile
  `Data Inizio`: 01.06.19
  `Durata`: 1 anno

- `Titolo`: Machine learning algorithms to handle label imbalance in biomedical taxonomies
  `Ente finanziatore`: UNIMI - Piano di Sostegno alla Ricerca 2015-2017
  `Ruolo`: responsabile
  `Data inizio`: 01.01.18
  `Durata`: 1 anno

Ha partecipato/partecipa ai seguenti progetti:

- `Titolo`: Machine Learning and Big Data Analysis for Bioinformatics
  `Ente finanziatore`: UNIMI - Piano di Sostegno alla Ricerca 2015-2017 - Linea 2
  `Ruolo`: collaboratore
  `Durata`: 1 anno

- `Titolo`: HyperGeV: Detection of Deleterious Genetic Variation through Hyper-ensemble Methods.
  `Ente finanziatore`: CINECA
  `Ruolo`: collaboratore
  `Data inizio`: 01.01.17
  `Durata`: 1 anno

- `Titolo`: A predictive model or response to FINGOLIMOD: integration of clinics, neuroradiology and genetics.
  `Ente finanziatore`: Fondazione Centro San Raffaele Milano
  `Ruolo`: membro unità DI (due membri)
  `Data inizio`: 01.05.16
  `Durata`: 5 mesi

- `Titolo`: Discovering Patterns in Multi-Dimensional Data
  `Ente finanziatore`: UNIMI - Piano di Sostegno alla Ricerca 2015-2017 - Linea 2B
  `Ruolo`: collaboratore
  `Data inizio`: 01.16
  `Durata`: 1 anno

- `Titolo`: Graph-based methodologies for the automated inference in bio-medical ontologies
  `Ente finanziatore`: UNIMI - Piano di Sostegno alla Ricerca 2015-2017 - Linea 2A
  `Ruolo`: unico membro oltre al coordinatore
  `Data inizio`: 01.16
  `Durata`: 1 anno

- `Titolo`: Sviluppo di algoritmi e software per l'analisi dello stato della cromatina in genomi di eucarioti
  `Ente finanziatore`: Consiglio Nazionale delle Ricerche (CNR) - Progetto bandiera EPIGEN
  `Ruolo`: collaboratore
  `Data inizio`: 05.12
  `Durata`: 10 mesi

Inoltre, ha presentato i seguenti progetti ricerca:

- European Molecular Biology Organization (EMBO) Short Term Fellowships, 2016, sulle tematiche del progetto "Graph-based methodologies for the automated inference in bio-medical ontologies"

- Progetto DaaD Short-Term Grants, 2016. Titolo:"Imbalance-aware disease gene prioritization through heterogeneous data integration and disease similarities".

- Progetto DaaD Short-Term Grants, 2015. Titolo:"Imbalance-aware methodologies to exploit CTD DiseaseComp features for disease gene prioritization".

5 Pubblicazioni e Indicatori Bibliometrici Output Ricerca in Banche Dati Internazionali

È autore di **17 lavori** pubblicati su riviste internazionali e **19 lavori** presentati a convegni nazionali e internazionali, in aggiunta a **5 lavori** presentati a conferenze senza proceedings e **3 lavori** attualmente sottomessi a riviste internazionali.

Di seguito i principali indici bibliometrici:

- GOOGLE SCHOLAR (07.19)
  (`https://scholar.google.it/citations?user=eduk3B8AAAAJ&hl=en`)
  Numero documenti indicizzati: 38
  N. citazioni: 287
  Indice h: 9

- ELSEVIER SCOPUS (`www.scopus.com`, 07.19).
  Author ID: 36835331200
  Numero documenti indicizzati: 27
  Periodo di riferimento: 01.2010-07.2019
  N. citazioni: 154.
  Indice h: 8

6 Attività di Ricerca

La sua attività di ricerca ha prevalentemente riguardato l'analisi e lo sviluppo di nuovi modelli di apprendimento automatico, la cui applicazione ha interessato ambiti legati alla biologia computazionale e alla bio-medicina. La formazione di base in informatica e il forte interesse per le discipline e i problemi bio-medici, hanno favorito l'esplorazione di linee ricerca caratterizzate dalla sinergia tra apprendimento automatico e bioinformatica. Ha contribuito a consolidare l'applicazione delle reti neurali a problemi di classificazione o di ranking con l'introduzione di nuovi modelli di reti di Hopfield parametriche, le cui implementazioni ha reso pubblicamente fruibili mediante specifiche librerie software.

In particolare, la sua attenzione è stata rivolta allo studio di metodi di classificazione cost-sensitive, cioè in grado di giungere a inferenze non banali, come molti dei classificatori cost-insensitive, in presenza di dati altamente sbilanciati. Molti problemi reali, infatti, sono caratterizzati da notevole sbilanciamento delle etichette (es. malato/sano, positivo/negativo etc.): ne sono esempi l'individuazione delle e-mail SPAM, l'inferenza di raccomandazioni commerciali user-based, la classificazione di fenotipi umani anomali (vedi per esempio il database Human Phenotype Ontology). Queste tematiche di ricerca hanno trovato una utile applicazione in diversi problemi medici e biologici, quali la determinazione dei geni causa di patologie genetiche, la predizione della funzione bio-molecolare delle proteine, la riqualificazione applicativa dei farmaci esistenti.

Lavori attualmente in corso, con risultati parziali che hanno già prodotto alcune pubblicazioni, stanno inoltre riguardando un problema ancora più generale, quello cioè della classificazione gerarchica multi-etichetta, dove ogni istanza può possedere più etichette tra loro organizzate in rapporti di specializzazione. Tali studi coinvolgono modelli di label propagation multi-task, modelli di teoria dei giochi e lo sviluppo di nuovi modelli di reti Hopfield gerarchiche. Quest'ultimo argomento è particolarmente interessante e inesplorato. Non esistono infatti in letteratura approcci simili, e molti sono i punti su cui investigare approfonditamente, quali ad esempio il trasferimento di informazione tra i diversi task e la convergenza della dinamica delle rete.

Infine, ha sviluppato e reso pubblicamente accessibili diversi pacchetti e librerie software relative ai metodi sviluppati.

All'interno del suo lavoro di ricerca si possono distinguere due aree principali, ulteriormente suddivise in sottoaree: apprendimento automatico e bioinformatica.

6.1 Apprendimento automatico

Riguardo la sua ricerca in apprendimento automatico, possiamo individuare le seguenti sottoaree:

6.1.1 Sviluppo di classificatori binari per dati altamente sbilanciati basati su reti di Hopfield parametriche

La linea principale di ricerca riguarda il problema della predizione di etichette binarie dei nodi in grafi parzialmente etichettati, che è un problema generale ricorrente molto di frequente nei casi reali. Le connessioni del grafo rappresentano similarità precomputate tra nodi, mentre le etichette dei nodi rappresentano la loro appartenenza (nodi positivi) o non appartenenza (nodi negativi) ad una determinata classe (proprietà o caratteristica specifica dei nodi). Inoltre, i nodi sono solo parzialmente etichettati, e l'obiettivo è estendere in maniera accurata la classificazione alla parte non nota del grafo.

In molti casi reali purtroppo e per varie ragioni, le etichettature sono spesso fortemente sbilanciate verso i negativi, e questo è un problema serio che determina un notevole decadimento di prestazioni, se non affrontato efficacemente. Infatti, la maggior parte dei metodi di predizione, in tali casi, tenderebbe a classificare negativamente tutte o quasi tutte le istanze.

Gran parte del suo lavoro di ricerca è stato dedicato allo sviluppo di classificatori *cost-sensitive*, cioè classificatori in grado proprio di gestire automaticamente l'elevato sbilanciamento delle etichette. In particolare, si è dedicato allo studio di un nuovo modello di rete di Hopfield discreto e parametrico, *COSNet* (**CO**st-**S**ensitive Neural **Net**work), in cui i parametri da apprendere sono le soglie dei neuroni e i loro valori di attivazione. Il modello infatti, separa concettualmente le etichette (di solito i positivi sono indicati con 1 e i negativi con -1 o 0) e i valori di attivazione (corrispondenti agli stati acceso / spento), permettendo di apprendere i valori di attivazione in modo da 'controbilanciare' la predominanza di neuroni negativi. I positivi, pertanto, possono propagare le loro etichette durante la dinamica della rete. Il metodo è stato confrontato con lo stato dell'arte in problemi complessi multi-classe multi-etichetta, come il problema della predizione della funzione e della espressione genica, con risultati promettenti [18, 37]. Ha studiato anche una versione regolarizzata del metodo per gestire diversi casi limite dell'algoritmo di apprendimento [19].

Inoltre, il modello di rete di Hopfield sviluppato è stato anche adottato per fornire un ranking dei nodi rispetto alla classe da predire, cioè un ordinamento dei nodi secondo il loro maggiore o minore potenziale appartenenza alla classe da predire [16, 35].

Tale modello è stato poi anche ripensato per una sua esecuzione in parallelo su architetture GPU, sfruttando la possibilità di aggiornare in parallelo sottoinsiemi di nodi (tra loro indipendenti) mediante graph coloring, pur mantenendo asincrona la dinamica generale della rete, condizione necessaria per la convergenza della dinamica stessa ad un punto fisso. Sfruttando anche una rappresentazione sparsa della rete, il nuovo modello, ParCOSNet, è stato applicato con bassi tempi di esecuzione a grafi con milioni di nodi utilizzando computer con harware standard, con l'unico requisito di essere equipaggiati di dispositivo GPU [5].

Di recente, ha sviluppato un nuovo modello di rete di Hopfield parametrica multi-categoria, *HoMCat* (**Ho**pfield **M**ulti-**Cat**egory) [14]. Il modello prevede categorie multiple di neuroni, determinate indipendentemente dalla classe da predire. Tale approccio deriva dalla constatazione che in diversi contesti reali le istanze da predire sono naturalmente partizionabili in diverse categorie secondo proprietà intrinseche alle istanze stesse. Il criterio di partizione in categorie è quindi diverso da quello individuato dalla classe funzionale che si vuole predire. A ogni categoria viene assegnato un parametro dedicato, da apprendere in modo da cogliere al meglio la struttura topologica della

categoria nel grafo. L'algoritmo di apprendimento adottato è molto efficiente e la sua complessità cresce solo linearmente con il numero di categorie. Due differenti modelli sono stati sviluppati: uno che apprende la partizione in categorie dei nodi di input automaticamente dai dati [36] e uno che invece la riceve come ulteriore input [14].

Il metodo è stato applicato con risultati molto positivi al problema predizione multi-specie della funzione proteica, in cui il grafo è composto di geni/nodi appartenenti a specie diverse ma collegate filogenicamente (es. topo e ratto). In tale contesto, ogni categoria rappresenta una diversa specie di cui si vogliono predire le funzioni molecolari, caratteristica difatti intrinseca ai nodi [14].

6.1.2 Algoritmi di apprendimento automatico per l'integrazione di dati eterogenei

È ben noto che le soluzioni a problemi di classificazione basati su grafi sono fortemente dipendenti dalla topologia del grafo stesso. In diversi contesti, le sorgenti di informazione che descrivono le relazioni di similarità tra istanze, tradotte poi in connessioni del grafo, sono diverse e variegate. Poiché ogni sorgente di dati cattura solo alcune delle caratteristiche funzionali dei nodi del grafo, singole sorgenti di dati possono essere in genere predittive per alcune classi (ogni classe rappresenta un problema di classificazione distinto), mentre possono risultare totalmente non informative per altre. Per questa ragione, l'integrazione di diverse sorgenti di dati in un unico grafo più affidabile e informativo, oltreché a più alta copertura, è un problema centrale in diversi contesti, quali ad esempio la bioinformatica.

Molti dei metodi presenti in letteratura per l'integrazione di diversi grafi in un grafo "consensus", tuttavia, trascurano completamente uno dei problemi principali che affliggono diversi contesti reali, e cioè lo sbilanciamento marcato delle etichette verso gli esempi negativi. Tale caratteristica del problema di classificazione invece risulta fondamentale anche nel decidere l'informatività di ogni singolo grafo.

In questo contesto, il candidato ha sviluppato e pubblicato sulla rivista principale di biologia computazionale [15] un nuovo metodo per l'integrazione di grafi eterogenei, *UNIPred* (**Un**balance-aware **I**ntegration and **Pred**iction), che tiene esplicitamente in considerazione lo sbilanciamento dei dati nel definire il livello di informatività di un singolo grafo relativamente a ogni singola classe. In particolare, agendo su un grafo per volta, il metodo introduce una proiezione dei nodi etichettati del grafo nel piano, cosicché la posizione dei punti proiettati sia rappresentativa dello sbilanciamento delle etichette nel rispettivo vicinato [33, 38]. Viene quindi introdotta per i punti proiettati una funzione obiettivo sensibile allo sbilanciamento, la cui ottimizzazione fornisce un indice di informatività del grafo per la classe da predire. L'ottimizzazione viene realizzata mediante un efficiente algoritmo approssimato, che rende agevole l'applicazione del metodo a grafi di grandi dimensioni.

Ha lavorato inoltre alla realizzazione di un servizio web che implementi l'algoritmo e lo renda pubblicamente fruibile, con una interfaccia user-friendly che ne permetta l'utilizzo anche a chi non ha familiarità con la programmazione. Lavoro sottomesso a rivista internazionale di punta del settore [4]. Infine, una estensione del modello è stata applicata con successo anche al problema della individuazione dei geni associati a patologie genetiche [25].

6.1.3 Selezione di esempi negativi in problemi di classificazione Positive-Unlabeled

Gli esempi negativi, che sono richiesti dalla maggioranza degli algoritmi di apprendimento automatico per inferire nuova conoscenza, soltanto raramente sono direttamente definiti e annotati in database pubblici. Infatti, si parla spesso di problemi Positive-Unlabeled (PU), cioè di problemi in cui solo le istanze positive per una determinata classe sono note, mentre le altre sono considerate non etichettate.

A tal fine, ha studiato e sviluppato un algoritmo per la selezione dei negativi più affidabili, basato su reti di Hopfield, che sfrutta una naturale clusterizzazione dei nodi positivi. L'algoritmo

è stato validato con risultati promettenti in un problema PU reale [31]. Sta inoltre lavorando a un modello di selezione delle istanze negative attraverso il paradigma dell'*active learning*, o apprendimento 'attivo', dove è il classificatore stesso che decide quali siano le istanze da selezionare come negative. Tale studio ha prodotto risultati molto competitivi parzialmente sottomessi [23], mentre è in fase avanzata un lavoro in via di sottomissione a rivista.

6.1.4 Sviluppo e analisi di metodi multitask e multietichetta per problemi di classificazione gerarchica

Questa parte della sua ricerca riguarda problemi di classificazione in cui più task vengono appresi contemporaneamente (metodi *multitask*), dove un nodo può appartenere a più classi al contempo (problemi *multietichetta*) e in cui i task sono gerarchicamente in relazione tra loro. I task sono dell'ordine delle centinaia o delle migliaia, a seconda del problema trattato. I task della gerarchia risultano più simili ad alcuni task che ad altri, e apprendere task simili tra loro al contempo in genere può migliorarne l'apprendimento rispetto al caso in cui i task vengano appresi singolarmente.

Ha sviluppato il primo modello di rete di Hopfield multitask, capace tra l'altro di classificare dati sbilanciati [21]. Più task 'simili' vengono appresi simultaneamente da un'unica rete di Hopfield, che apprende i parametri del modello in modo da ottimizzare un criterio di loss multitask globale, mentre le nuove etichette vengono predette mediante l'esecuzione della dinamica della rete fino al raggiungimento di uno stato di equilibrio.

L'approccio tradizionale dei metodi multitask tende ad apprendere modelli predittivi per ogni task in modo da rendere 'più vicine' le etichettature assegnate a task simili. Di recente si è dedicato a esplorare anche la possibilità che metodi multitask riescano a migliorare la classificazione imparando anche da task *dissimili*. In particolare, ha studiato metodi semi-supervisionati di label propagation multitask basati su diverse misure di dissimilarità tra i task. Si è provato che questo approccio, che tende ad assegnare etichettatture più *distanti* a task dissimili, è particolarmente utile in casi in cui le etichettature dei task siano sbilanciate verso i negativi, e applicazioni a casi reali hanno confermato i risultati teorici [29]. Una estensione di questo lavoro applicata al problema della predizione della funzione proteica è stata pubblicata su una rivista internazionale [10].

Ha proposto inoltre un altro interessante approccio allo stesso problema nel lavoro [8], in cui è stata proposta una metodologia di classificazione multitask basata su grafi e modelli di teoria dei giochi, dove ogni proteina (vertice del grafo) è un giocatore e le funzioni da predire (task) sono le strategie di gioco. La similarità tra task viene inglobata in una opportuna funzione di 'payoff', e il concetto di equilibrio di Nash permette di assegnare etichettature già rispettose della struttura gerarchica. Le prestazione del metodo in un problema reale multitask sono risultate a livello dei migliori metodi in letteratura [8].

Ancora in tale contesto, un altro problema e rappresentato dal fatto che algoritmi di apprendimento che classificano le istanze per ciascuna classe indipendentemente dalle altre, spesso violano i reapporti gerarchici tra le classi padre e le classi discendenti (i rapporti padre-figlio sono specializzazioni: una istanza non può essere al contempo positiva per una classe figlio e negativa per quella padre). Sono diversi i contesti reali in cui le classi sono strutturate come grafi diretti aciclici (DAG), come la Gene Ontology e la Human Phenotype Ontology.

Un utile strumento in questi casi è rappresentato dai metodi di correzione gerarchica post-predizione, che opportunamente rimodulano le predizioni fornite dai singoli classificatori per ottenere predizioni che siano coerenti con i vincoli di specializzazione padre-figlio tra le classi.

A tal proposito, ha contributo a sviluppare un metodo di correzione gerarchica per classi organizzate in DAG che si basa su due passi principali: una propagazione delle predizioni positive dalle foglie alle radici del DAG (bottom-up), seguita da una scansione inversa (top-down) del DAG per ripristinare i vincoli padre-figlio. Essendo presenti in un DAG potenzialmente più cammini da una radice a un nodo, per garantire che ogni nodo venga elaborato una sola volta, nel passo top-down i nodi vengono scanditi per livello, dove i livelli sono definiti come distanza massima dei nodi dalla

radice. I risultati preliminari ottenuti sono incoraggianti e sono stati pubblicati in due conferenze internazionali [26, 32].

Nello stesso contesto, cioè quello della classificazione gerarchica con metodi multi-classe multi-etichetta, di notevole importanza è l'individuazione di misure affidabili che descrivano il livello di similarità tra classi distinte, da usare per esempio in metodi di *Transfer Learning*. A tal fine, in collaborazione con il Computer Science and Centre for Systems and Synthetic Biology della Royal Holloway di Londra, ha contribuito a sviluppare un web-server *GOssTo* (the **G**ene **O**ntology **S**emantic **S**imilarity **To**ol) per la computazione di misure di similarità semantica tra classi della gerarchia Gene Ontology, che descrive gerarchicamente mediante un DAG le relazioni tra funzioni bio-molecolari dei geni [17, 34]. Il metodo è disponibile sia come web server che come applicazione JAVA standalone[19].

6.1.5 Sviluppo di librerie software di machine learning

Ha sempre direttamente implementato gli algoritmi sviluppati e utilizzati, resi poi disponibili su repository pubblici, quali *GitHub* e *Bioconductor*, nonnché mediante Original Software Publication. Di seguito l'elenco:

- **COSNet**. Package **R** scaricabile da Bioconductor[20], principale repository di Bioinformatica, e dal repository software GitHub[21], implementa l'algoritmo di classificazione omonimo [18], insieme a diverse utility ausiliarie, tra cui quelle per valutare le capacità di generalizzazione dell'algoritmo mediante k-fold cross validation. L'implementazione si serve del linguaggio **R** per le funzioni di routine, mentre, per ragioni di efficienza, le parti computazionalmente pesanti sono scritte in linguaggio **C**. *COSNet* è ora disponibile anche come *Original Sofware Publication* sulla rivista internazionale *Neurocomputing* [13]. Il software sta ricevendo più di un migliaio di download distinti all'anno da quando è stato pubblicato[22].
- **UNIPred**. Libreria, scritta nei linguaggi **C** ed **R**, che implementa l'omonimo metodo [15]. Il software è pubblicamente accessibile all'indirizzo `http://frasca.di.unimi.it/Unipred.html`.
- **GOssTo**. Applicazione JAVA standalone e web-server[23] per la computazione di similarità semantica tra classi funzionali della Gene Ontology (GO). Il servizio è molto flessibile, e permette di selezionare, oltre alle classi di interesse, anche l'organismo su cui operare e, al suo interno, i geni di interesse. Inoltre, l'applicazione fornisce anche la possibilità di calcolare la similarità tra geni in base alle loro annotazioni GO.
- **RANKS**. Application note della rivista Bioinformatics che descrive il package **R** *RANKS* [12]. Il package, disponibile sul repository pubblico CRAN (`http://cran.r-project.org`), implementa in linguaggio **C** ed **R**, un metodo di ranking di nodi in grafi basato su funzioni kernel. Il package fornisce anche l'implementazione di altri metodi dello stato dell'arte, e utility per il confronto delle prestazioni dei diversi metodi.
- **ParCOSNet**. **C++** codice sorgente ed eseguibile, pubblicamente accessibile dal repository GitHub[24], che implementa l'algoritmo ParCOSNet [5], versione parallela e sparsa dell'algoritmo COSNet e capace di predire le etichette binarie dei nodi di grafi sparsi, aventi milioni di nodi, anche su macchine standard, purché dotate di architetture GPU.

[19] `http://www.paccanarolab.org/gossto`
[20] `https://www.bioconductor.org/packages/release/bioc/html/COSNet.html`
[21] `https://github.com/m1frasca/COSNet_GitHub`
[22] `http://bioconductor.org/packages/stats/bioc/COSNet/`
[23] `http://www.paccanarolab.org/gossto/`
[24] `https://github.com/anacleto63/ParCOSNet`

6.2 Bioinformatica

Le attività di ricerca in bioinformatica del candidato hanno riguardato lo sviluppo e l'applicazione di metodi e algoritmi di apprendimento automatico per l'estrazione di conoscenza biologica dall'analisi di dati bio-molecolari generati da bio-tecnologie high-throughput. Possiamo distinguere le seguenti linee di ricerca:

6.2.1. Classificazione funzionale di geni e proteine

6.2.2. Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche

6.2.3. Metodi per la predizione di associazioni gene-patologia per le principali patologie Mendeliane

6.2.4. Predizione multi specie della funzione di geni e proteine

6.2.5. Analisi epigenetica dei livelli di esperessione genica

6.2.6. Generazione di dati biologicamente motivati per la valutazione di metodi di selezione dei geni associati a fenotipi patologici

6.2.7. Sviluppo di metodi per il riposizionamento di farmaci in nuove classi terapeutiche

6.2.1 Classificazione funzionale di geni e proteine

La classificazione funzionale dei numerosi geni e proteine appartenenti a diversi organismi sequeziati a partire dagli anni 2000 è diventata un problema centrale in bioinformatica. Le varie funzioni biomolecolari che le proteine possono possedere sono strutturate mediante rigide gerarchie che ne descrivono i rapporti di specializzazione: esempi sono la Gene Ontology (GO), un grafo diretto aciclico, e gli alberi FunCat del MIPS di Monaco. Il problema è molto complesso, e di molte delle migliaia di proteine sequenziate ancora non si possiedono annotazioni (associazioni delle proteine con termini della gerarchia). Inoltre, le proteine possono appartenere a più termini della ontologia (problema multi-label) e per ogni singola funzione biologica, spesso pochissime annotazioni sono note (sbilanciamento verso i negativi). Infine, recentemente l'analisi di dati biologici rappresentati sotto forma di grafo ha acquisito sempre maggior popolarità. Tale trend è facilmente motivabile considerando che le funzioni biologiche sono estremamente complesse e, di conseguenza, solo di rado avviene che una determinata funzione sia sotto il diretto controllo di un unico gene o di un'unica proteina. Le reti biologiche costituiscono quindi il mezzo di elezione per l'analisi delle complesse interazioni tra geni che, in caso di alterazione, possono anche essere all'origine di importanti patologie. Pertando si richedono metodi in grado di gestire tale rappresentazione.

A tal fine, la sua attività di recerca degli ultimi anni ha riguardato lo sviluppo di classificatori basati su reti biolgiche per geni e proteine di organismi modello e per i termini delle principali gerarchie funzionali [12, 13, 16, 19]. In particolare, ha applicato i modelli cost-sensitive sviluppati in [18, 37] con risultati decisamente competitivi con lo stato dell'arte per la predizione della funzione genica.

Nello stesso contesto, ricordando che nelle gerarchie funzionali descritte le annotazioni delle proteine per le classi figlio vengono trasferite anche alle classi padre, metodi di predizione della funzione proteica che considerano un solo termine della gerarchia per volta, possono violare questo vincolo padre-figlio. A questo scopo, il metodo di correzione gerarchica post predizione sviluppato in [26, 32] è stato applicato per ripristinare tali vincoli nelle gerarchie Gene Ontology e Human Phenotype Ontology, con risultati preliminari che hanno già mostrato l'efficacia del metodo.

Inoltre, nello stesso contesto gerarchico, esistono anche approcci che cercano di predire più funzioni bio-molecolari per volta, sfruttandone la correlazione e le relazioni di similarità. Per permettere a questi metodi di aggregare tra loro funzioni simili, ha sviluppato in collaborazione con il

Computer Science and Centre for Systems and Synthetic Biology della Royal Holloway di Londra, il già menzionato web-server (*GOssTo*) per la computazione di diverse misure di similarità semantica tra termini della gerarchia Gene Ontology [17, 34].

Individuazione delle proteine negative. Infine, di recente la sua attività di ricerca ha riguardato anche il problema della individuazione delle proteine considerate negative per un determinato termine GO. Infatti, l'ontologia GO memorizza in genere soltanto gli esempi positivi (proteina associata al termine) e tutte le proteine non positive in principio potrebbero essere considerate come esempi negativi per addestrare dei classificatori binari per predire le associazioni tra le proteine e il termine stesso. In questo contesto, ha anzitutto condotto un'analisi aprofondita delle nuove annotazioni proteina-funzione proteica ricevute in un arco temporale di due anni e relative a due diverse versioni temporali della GO, al fine di caratterizzare il collocamento di tali annotazioni nel DAG e la loro relazione con le annotazioni esistenti. Questo studio ha rivelato informazioni molto interessanti, che, in uno studio ancora preliminare, hanno condotto allo sviluppo di un nuovo metodo di selezione dei negativi per funzioni GO, con risultati competitivi con lo stato dell'arte [24]. Inoltre, sono stati condotti studi per l'individuazione di feature proteiche estratte dal grafo delle proteine e caratterizzanti il singolo nodo (proteina), con un'accurata analisi di quali di queste feature fossero più rilevanti per il problema della selezione dei negativi, includendo più di una dozzina di features tra quelle proposte in letteratura e di carattere topologico sia locale (vicinato della proteina) che globale (tutto il grafo) [7, 27]. Nello specifico, sono state prese in considerazione le principali misure di centralità in grafi, incluse le corrispondenti versioni *function-dependent*, di cui si è validata la loro rilevanza sia nella selezione dei negativi, che nel predire le funzioni GO stesse.

6.2.2 Integrazione di sorgenti multiple di dati per la predizione delle funzioni geniche

Altro aspetto rilevante e complesso del problema della predizione funzione genica è la presenza di diverse sorgenti dati (co-espressione genica, famiglie proteiche, dati di sequenza, interazioni genetiche, etc.), ciascuna delle quali reappresentabile mediante una rete di geni a diversa copertura e informatività. Siccome ogni sorgente di dati cattura solo alcune delle caratteristiche funzionali dei geni e dei loro prodotti, singole sorgenti di dati biomolecolari sono in genere predittive solo per alcune classi funzionali, mentre possono risultare totalmente non informative per altre. A tal fine, ha applicato il metodo sviluppato in [15] nella integrazione di reti genetiche di diversi organismi, tra cui uomo, moscerino, lievito, e in particolare topo, dove si è confrontato con buoni risultati con i metodi della challange internazionale MOUSEFUNC I.

6.2.3 Metodi per la predizione di associazioni gene-patologia per le principali patologie Mendeliane

Altro problema centrale in bioinformatica è la determinazione dei geni associati alle patologie Mendeliane e alle patologie genetiche complesse. Infatti, raramente una patologia genetica è causata da un solo gene, ma spesso è dovuta alla azione combinata di pathway genetici. Individuare manualmente i geni causali, oltreché costoso, è in sostanza impraticabile, dato l'elevato numero di geni umani (più di 20000) e di patotologie (migliaia). Pertanto si sono resi necessari metodi automatici per fornire degli indizi su quali geni potessero essere candidati per specifiche patologie.

In questo contesto, ha sviluppato un metodo veloce e scalabile per determinare un ranking dei geni per una data patologia: più alta la posizione nel ranking, più forte l'indizio che il gene sia tra le cause della patologia. Il metodo è risultato primo tra i metodi di un benchmark pubblico preparato per testare l'efficacia dei metodi automatici per l'individuazione di geni candidati per circa 700 patologie catalogate nel Medical Subject Headings (MeSH) database [30]. Una prima estensione

del modello, che include l'applicazione di un nuovo metodo per la selezione dei negativi in fase di apprendimento, ne ha ulteriormente migliorato le prestazioni [11]. Infine, la nuova idea di sfruttare le similarità dei profili genetici tra patologie è stata inclusa nel metodo, ottendo una metodologia in grado di raggiungere accuratezze elevatissime nella prioritizzazione dei geni patologici e di fornire, unico allo stato dell'arte, un ordinamento dei geni anche per le patologie genetiche per cui ancora non esistono geni associati [9].

6.2.4 Predizione multi specie della funzione di geni e proteine

Grazie all'utilizzo di moderne biotecnologie, un singolo esperimento può produrre informazioni su milioni di molecole, e nell'ultima release della banca dati di riferimento delle proteine è possibile trovare informazioni su milioni di proteine provenienti da migliaia di specie differenti. Alcune specie sono tra di loro correlate filogeneticamente, cioè possiedono antenati comuni nel processo evolutivo. Questo vuol dire che proteine omologhe appartenenti a specie diverse possono svolgere funzioni simili. Pertanto, integrare in un'unica rete proteine appertenenti a specie filogeneticamente vicine può migliorare la predizione su ciascuna specie.

La predizione multi-specie è uno dei contesti appropriati per l'applicazione del metodo sviluppato in [15], dove le categorie sono proprio le singole specie. I risultati ottenuti hanno mostrato un notevole incremento di prestazioni rispetto alle reti con singola categoria, ed è in corso un lavoro estensivo per applicare tale metodo a numerose reti contenenti centinaia di specie.

Inoltre, le difficoltà della predizione multi-specie riguardano anche le necessità di integrare sorgenti multiple e di visualizzare in maniera approriata i risultati di tale predizione. A tal fine, di recente ha pubblicato un lavoro in cui viene descritto e proposto un framework che connette i momenti principali della predizione multi-specie: recupero ed elaborazione di dati biomolecolari da diverse fonti e diversi organismi, relativa integrazione in un'unica grande rete, predizione delle funzioni biomolecolari e visualizzazione dei risultati [28].

6.2.5 Analisi epigenetica dei livelli di esperessione genica

. L'espressione genica è un processo molto complesso, finemente regolato sia a livello genetico che epigenetico. In particolare, quest'ultimo riguarda la cromatina, struttura proteica formata da istoni che avvolge il DNA. Di recente si è mostrato come la cromatina sia un fattore chiave nella determinazione della espressione genica. Quindi, costruire modelli che predicano i livelli di espressione genica a partire da fattori epigentici (modifiche istoniche, attività dei fattori di trascrizione, etc.) è un problema attuale in bioinformatica.

Nel suo periodo da assegnista presso il Dipartimento di Bioscienze, ha studiato un modello parametrico basato su reti neurali che ordina i geni in base all'informazione contenuta nelle modifiche istoniche nei pressi dei geni [35]. Il modello è stato validato nel predire i livelli di espressione genica di 6 linee cellulari umane, considerando tutto il genoma.

6.2.6 Generazione di dati biologicamente motivati per la valutazione di metodi di selezione dei geni associati a fenotipi patologici

La validazione dei metodi di selezione dei geni è un problema serio in bioinformatica. In molti casi, i geni associati a uno specifico fenotipo non sono noti a priori, e pertanto la valutazione della efficacia di metodi di selezione genica è difficile e solo parzialmente eseguita usando algoritmi di classificazione. In questo contesto, ha contribuito a sviluppare un modello matematico per generare dati di espressione genica ('quantità' di prodotti del gene sotto determinate condizioni) biologicamente plausibili, per testare i metodi di selezione genica. Il metodo modella i profili di espressione e le firme di espressione di un fenotipo specifico attraverso funzioni Booleane positive.

Date le firme di espressione e i geni associati con il fenotipo di interesse, il metodo è in grado di generare dati di espressione genica statisticamente plausibili dal punto di vista biologico [20].

6.2.7 Sviluppo di metodi per il riposizionamento di farmaci in nuove classi terapeutiche

Lo sviluppo di nuovi farmaci è un processo costoso e fortemente soggetto a possibili fallimenti. Negli ultimi anni un nuovo paradigma di ricerca farmacologica noto come "Drug Repositioning" (riposizionamento di farmaci in classi terapeutiche differenti da quelle per cui erano stati inizialmente sviluppati) sta emergendo in quanto è in grado di ridurre i costi di sviluppo e i tempi necessari all'immissione di nuovi farmaci sul mercato che richiedono, tipicamente, 10-15 anni e investimenti che superano il miliardo di dollari. Anche in questo caso, i farmaci possono essere organizzati in una rete di farmaci le cui connessioni racchiudono diverse informazioni sui farmaci stessi (interazione con proteine specifiche, geni o patologie su cui il farmaco ha effetto, etc.). Il metodo proposto in [13] è stato applicato anche a questo problema, su dei dati pubblicamente disponibili. I risultati, in confronto con metodi specificamente studiati per questo problema, sono stati molto competitivi.

Pubblicazioni

Lavori attualmente sottomessi a riviste e conferenze internazionali con peer-review

S.1

[1] N. Zhou et al.. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 2019.

S.2

[2] A. Petrini, M. Mesiti, M. Schubach, **M. Frasca**, D. Danis, M. Re, G. Grossi, L. Cappelletti, T. Castrignano', P.N. Robinson, G. Valentini. parSMURF, a High Performance Computing tool for the genome-wide detection of pathogenic variants. *GigaScience*, 2019. In second revision.

S.3

[3] J. Gliozzo, P. Perlasca, M. Mesiti, A. Petrini, E. Casiraghi, **M. Frasca**, G. Grossi, M. Re, A. Paccanaro, and G. Valentini. Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction. *Scientific Reports*, 2019. In second revision.

Riviste internazionali con peer-review

J.1

[4] P. Perlasca, **M. Frasca**, C.T. Ba, M. Notaro, A. Petrini, E. Casiraghi, G. Grossi, J. Ghilozzo, G. Valentini, M. Mesiti. UNIPred-Web: a Web Tool for the Integration and Visualization of Biomolecular Networks for Protein Function Prediction. *BMC Bioinformatics*, 2019, in press. Impact factor: 2.970

J.2

[5] **M. Frasca**, G. Grossi, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, and G. Valentini A GPU-based algorithm for fast node label learning in large and unbalanced biomolecular networks. *BMC Bioinformatics*, 19 (Suppl 10):353, 2018. ISSN:1471-2105. doi:10.1186/s12859-018-2301-4. Impact factor: 2.970

J.3

[6] E. Casiraghi, V. Huber, **M. Frasca**, M. Cossa, M. Tozzi, L. Rivoltini, B.E. Leone, A. Villa and B. Vergani A novel computational method for automatic segmentation, quantification and comparative analysis of immunohistochemically labeled tissue sections. *BMC Bioinformatics*, 19 (Suppl 10):357, 2018. ISSN:1471-2105. doi:10.1186/s12859-018-2302-3. Impact factor: 2.970

J.4

[7] P. Boldi, **M. Frasca** and D. Malchiodi. Evaluating the Impact of Topological Protein Features on the Negative Examples Selection. *BMC Bioinformatics*, 19 (Suppl 14):417, 2018. ISSN:1471-2105.
doi:10.1186/s12859-018-2385-x. Impact factor: 2.970

J.5

[8] S. Vascon, **M. Frasca**, R. Tripodi, G. Valentini and M. Pelillo. Protein Function Prediction as a Graph-Transduction Game. *Pattern Recognition Letters*, 2018. In press. ISSN:0167-8655. doi:10.1016/j.patrec.2018.04.002. Impact factor: 2.810

J.6

[9] **M. Frasca**. Gene2DisCo: Gene to Disease Using Disease Commonalities. *Artificial Intelligence in Medicine*, 82:34–46, 2017. ISSN:0933-3657.
doi:https://doi.org/10.1016/j.artmed.2017.08.001. Impact factor: 3.574

J.7

[10] **M. Frasca** and N. Cesa Bianchi. Multitask protein function prediction through task dissimilarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017. ISSN:1545-5963.
In press. doi:10.1109/TCBB.2017.2684127. Impact factor : 2.896

J.8

[11] **M. Frasca** and D. Malchiodi. Exploiting negative sample selection for prioritizing candidate disease genes. *Genomics and Computational Biology*, 3(3) 47, 2017. ISSN:2365-7154.
doi:10.18547/gcb.2017.vol3.iss3.e47.

J.9

[12] G. Valentini, G. Armano, **M. Frasca**, J. Lin, M. Mesiti, and M. Re. RANKS: a flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, 32(18):2872–2874, 2016. ISSN 1367-4803.
doi: http://dx.doi.org/10.1093/bioinformatics/btw235. Impact factor : 4.531

J.10

[13] **M. Frasca** and G. Valentini. COSNet: an R package for label prediction in unbalanced biological networks. *Neurocomputing*, 2016. In press. Available online from 29 April 2016. ISSN 0925-2312.
doi: http://dx.doi.org/10.1016/j.neucom.2015.11.096. Impact factor : 4.072

J.11

[14] **M. Frasca**, S. Bassis, and G. Valentini. Learning node labels with multi-category Hopfield networks. *Neural Computing and Applications*, 27(6):1677–1692, 2016. ISSN 1433-3058.
doi: 10.1007/s00521-015-1965-1. Impact factor : 4.664

J.12

[15] **M. Frasca**, A. Bertoni, and G. Valentini. UNIPred: Unbalance-Aware Network Integration and Prediction of Protein Functions. *Journal of Computational Biology*, 22(12):1057–1074, 2015. ISSN 1066-5277.
doi: 10.1089/cmb.2014.0110. Impact factor : 1.191

J.13

[16] **M. Frasca**. Automated gene function prediction through gene multifunctionality in biological networks. *Neurocomputing*, 162:48 – 56, 2015. ISSN 0925-2312.
doi: 10.1016/j.neucom.2015.04.007. Impact factor : 4.072

J.14

[17] H. Caniza, A. E. Romero, S. Heron, H. Yang, A. Devoto, **M. Frasca**, M. Mesiti, G. Valentini, and A. Paccanaro. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics*, 30(15):2235–2236, 2014. ISSN 1367-4803.
doi: 10.1093/bioinformatics/btu144. Impact factor : 4.531

J.15

[18] **M. Frasca**, A. Bertoni, M. Re, and G. Valentini. A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, 43:84 – 98, 2013. ISSN 0893-6080.
doi: 10.1016/j.neunet.2013.01.021. Impact factor : 5.785

J.16

[19] **M. Frasca**, A. Bertoni, and G. Valentini. Regularized network-based algorithm for predicting gene functions with high-imbalanced data. *EMBnet.journal*, 18(15):41–42, 2012. ISSN 2226-6089. doi: 10.14806/ej.18.A.377

J.17

[20] M. Muselli, A. Bertoni, **M. Frasca**, A. Beghini, F. Ruffino, and G. Valentini. A mathematical model for the validation of gene selection methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(5):1385–1392, Sept 2011. ISSN 1545-5963. doi: 10.1109/TCBB.2010.83. Impact factor : 2.896

Atti di conferenze internazionali e nazionali

C.1

[21] **M. Frasca**, G. Grossi, G. Valentini. Multitask Hopfield Networks. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML19), 2019. Accepted.

C.2

[22] C. T. Ba, E. Casiraghi, **M. Frasca**, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, M. Re and G. Valentini. A Graphical Tool for the Exploration and Visual Analysis of Biomolecular Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. In press.

C.3

[23] **M. Frasca**, M. Sepehri, A. Petrini, G. Grossi and G. Valentini. Committee-based Active Learning to Select Negative Examples for Predicting Protein Functions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. In press.

C.4

[24] M. Sepehri, **M. Frasca**. Analysis of Novel Annotations in the Gene Ontology for Boosting the Selection of Negative Examples. *9th International Conference on Biomedical Engineering and Technology (ICBET)*, 2019, in press.

C.5

[25] **M. Frasca**, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi and M. Andrade-Navarro. Disease–Genes must Guide Data Source Integration in the Gene Prioritization Process. *Computational Intelligence Methods for Bioinfor- matics and Biostatistics, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 60–69, Springer, Cham, 2019. ISSN:0302-9743.

C.6

[26] M. Notaro, M. Schubach, **M. Frasca**, M. Mesiti, P.N. Robinson and G. Valentini. Ensembling Descendant Term Classifiers to Improve Gene–Abnormal Phenotype Prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, in press. ISSN:0302-9743.

C.7

[27] **M. Frasca**, F. Lipreri and D. Malchiodi. Analysis of Informative Features for Negative Selection in Protein Function Prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10209 LNCS, 267-276, 2017. doi: 10.1007/978-3-319-56154-7_25. ISSN:0302-9743.

C.8

[28] G. Perlasca, P. Valentini, **M. Frasca**, and M. Mesiti. Multi-species protein function prediction: Towards web-based visual analytics. In *18th International Conference on Information Integration and Web-based Applications & Services(iiWAS2016), 28–30 Nov, Singapore.*, pages 1–5. ACM 2016. ISBN 978-1-4503-4807-2.

C.9

[29] **M. Frasca** and N. Cesa Bianchi. Multi-task label propagation with dissimilarity measures. In *15th International Workshop on Data Mining in Bioinformatics (BIOKDD 16), 14 Aug, San Francisco, CA, USA.*, pages 1–8, 2016. URL : `http://home.biokdd.org/biokdd16/`.

C.10

[30] **M. Frasca** and S. Bassis. Gene-Disease Prioritization Through Cost-Sensitive Graph-Based Methodologies, volume 9656 of *Lecture Notes in Computer Science*, pages 739–751. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-31744-1_64.
ISBN 978-3-319-31744-1.

C.11

[31] **M. Frasca** and D. Malchiodi. Selection of negative examples for node label prediction through fuzzy clustering techniques. In *Advances in Neural Networks. WIRN 2015. Smart Innovation, Systems and Technologies*, pages , vol 54:67–76, Springer, Cham. doi: 10.1007/978-3-319-33747-0_7.
ISBN:978-3-319-33746-3.

C.12

[32] P. N. Robinson, **M. Frasca**, S. Köhler, M. Notaro, M. Re, and G. Valentini. *A Hierarchical Ensemble Method for DAG-Structured Taxonomies*, *Multiple Classifier Systems. MCS 2015. Lecture Notes in Computer Science*, vol 9132 pages 15–26, 2015. Springer International Publishing, Cham. doi: 10.1007/978-3-319-20248-8_2.
ISBN 978-3-319-20248-8.

C.13

[33] **M. Frasca**, A. Bertoni, and G. Valentini. An unbalance-aware network integration method for gene function prediction. In *MLSB 2013 - Machine Learning for Systems Biology - Berlin, July 19-20*, Berlin, Germany, 2013.

C.14

[34] H.V. Caniza, A.E. Romero, S. Heron, H. Yang, **M. Frasca**, M. Mesiti, G. Valentini, and A. Paccanaro. Gossto & gosstoweb: user-friendly tools for calculating semantic similarities on the gene ontology. In *Bio-Ontologies SIG 2013 - ISMB 2013*, Berlin, Germany, 2013.

C.15

[35] **M. Frasca** and G. Pavesi. A neural network based algorithm for gene expression prediction from chromatin structure. In *The 2013 International Joint Conference on Neural Networks (IJCNN), 4-9 Aug, Dallas Texas,*, pages 1–8. IEEE, 2013. doi: 10.1109/IJCNN.2013.6706954.
ISBN 978-1-4673-6128-6.

C.16

[36] **M. Frasca**, A. Bertoni, and A. Sion. A Neural Procedure for Gene Function Prediction. *Neural Nets and Surroundings. Smart Innovation, Systems and Technologies*, vol 19, pages 179–188, 2013. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-35467-0_19.
ISBN 978-3-642-35467-0.

C.17

[37] A. Bertoni, **M. Frasca**, and G. Valentini. COSNet: A Cost Sensitive Neural Network for Semi-supervised Learning in Graphs. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science*, vol 6911, pages 219–234, 2011. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-23780-5_24.
ISBN 978-3-642-23780-5.

C.18

[38] A. Bertoni, **M. Frasca**, and G. Valentini. An efficient supervised method to integrate multiple biological networks. In *BITS 2011, Bioinformatics Italian Society Annual Meeting, June 20-22*, Pisa, Italy, 2011.

C.19

[39] A. Bertoni, **M. Frasca**, G. Grossi, and G. Valentini. *Learning functional linkage networks with a cost-sensitive approach*, volume 226 of *Frontiers in Artificial Intelligence and Applications*, pages 52–61. IOS Press, 2010. doi: 10.3233/978-1-60750-692-8-52.

Presentazioni a conferenze senza Proceedings

C. T. Ba, E. Casiraghi, **M. Frasca**, J. Gliozzo, M. Mesiti, M. Notaro, P. Perlasca, A. Petrini, M. Re and G. Valentini. A Graphical Tool for the Exploration and Visual Analysis of Biomolecular Networks. Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB), 2018, 6-8 Settembre, Caparica, Portugal, 2018.

M. Frasca, M. Sepehri, A. Petrini, G. Grossi and G. Valentini. Committee-based Active Learning to Select Negative Examples for Predicting Protein Functions. *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB), 2018, 6-8 Settembre, Caparica, Portugal*, 2018.

M. Frasca, J.F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi and M. Andrade-Navarro. Disease–Genes must Guide Data Source Integration in the Gene Prioritization Process. *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, 2017, 7-9 Settembre, Cagliari, Italy.

M. Notaro, M. Schubach, **M. Frasca**, M. Mesiti, P.N. Robinson and G. Valentini. Ensembling Descendant Term Classifiers to Improve Gene–Abnormal Phenotype Prediction. *Computational Intelligence methods for Bioinformatics and Biostatistics (CIBB)*, 2017, 7-9 settembre, Cagliari, Italy.

M. Frasca. *Selection of Negatives in Hopfield Networks*, *International Workshop on Dynamics of Multi-Level Systems (DYMULT) 2015, Max Planck Institute for the Physics of Complex Systems*, Dresden, 2015.