ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI
FARMACIA E BIOTECNOLOGIE

Oggetto: Piano formativo assegno BIR 2022 cofinanziato fondi Emidio Capriotti (CUP J33C21000210005)

Per ogni obiettivo del progetto di ricerca, Il piano formativo relativo all'assegno BIR 2022 cofinanziato al 50% con i fondi EMBL-ELIXIR prevede le seguenti le attività formative:

1. **Raccolta dati per la valutazione dei metodi per l'analisi del genoma umano**
   Nella prima fase del progetto (mesi 1-4), l'assegnista acquisirà la capacita di interrogare automaticamente database disponibili online tramite l'utilizzo di programmi. In particolare verranno estratte informazioni dalla banca dati ClinVar e integrati con quelle strutturali e funzionali disponibili sulle banche dati PDB e UniProt. Questi dati verranno combinati con le annotazioni dei genotipi delle diverse malattie rare di origine genetica riportate nella banca data Orphanet. Per costruire una rete di geni associati ad ogni classe di malattia verranno considerati dati di interazioni proteina-proteina. Il risultato di questa prima fase dell'attività consiste nello sviluppo di una collezione di mutazioni di singoli amino acidi e relativi geni associati a malattie rare e le reti geniche coinvolte.

2. **Valutazione e ottimizzazione dei metodi per la predizione i varianti genetiche patogene associate a malattie rare**
   Dal mese 5 al mese 8, l'assegnista istallerà e utilizzerà su infrastrutture computazionali, tramite accesso remoto, i metodi classici per la predizione dell'impatto delle mutazioni amino ecidiche sulla stabilità (FOLDX, DDGun) e sulla funzione (CADD, FATHMM, PhD-SNPg) delle proteine. Una volta ottenute le predizioni, dal mese 9 al mese 12, l'assegnista analizzerà le predizioni per identificare le soglie che permettono di ottimizzare le predizioni dei diversi metodi su specifiche classi di malattia. In particolare il lavoro si focalizzerà su dati di sequenziamento relativi a malattie del neuro sviluppo depositati nella banca dati EGA. Questi dati verranno confrontati con quelli messi a disposizione dal consorzio *1000 Genomes*.

3. **Sviluppo di metodi e piattaforme computazionale per la predizione del rischio di sviluppare malattie rare**
   Per raggiungere l'obbiettivo finale del progetto, dal mese 13 al 18, l'assegnista utilizzerà i dati analizzati durante il primo anno del progetto per sviluppare metodi probabilistici capaci di valutare per ogni individuo il rischio di sviluppare una malattia del neuro sviluppo a partire dalle mutazioni identificate nell'esoma. Nella fase finale del progetto, dal mese 19 al 24, l'assegnista svilupperà una piattaforma computazionale che permetta l'analisi automatica dell'esoma umano. Tale piattaforma sarà resa disponibile in modalità "*open access*" e testata anche in ambienti computazionali protetti dalla normativa sulla privacy.

   Nell'ambito del progetto l'assegnista avrà la possibilità di collaborare con ricercatori del Laboratorio di Genetica Medica dell'ospedale Sant'Orsola di Bologna e con ricercatori del nodo locale dell'INFN per raggiungere gli obiettivi 2 e 3 del progetto.

Emidio Capriotti
Via Selmi 3 | 40126 Bologna | Italia | Tel. + 39 051 2094303 | emidio.capriotti@unibo.it

Research Proposal - BIR 2022

Title: Computational infrastructure for the analysis of the rare disease genome
Supervisor: Emidio Capriotti
Duration: 24 months
Funding request: 24,000 Euro
Co-funding: 24,000 Euro (CUP: J33C21000210005)

Background

According to the definition adopted by the European Community, a rare disease (RD) affects fewer than 1 in 2,000 individuals. A recent analysis of the Orphanet database identified ~5,300 distinct types of RDs that affect between 3-6% of the population of the European Union (1). Although rare variants in ~4,000 genes have now been shown to be related to RDs, many more genes remain to be discovered. The recent evolution of sequencing technologies has revealed in several cases the molecular causes of single RDs and has allowed the identification of new missense mutations responsible for specific disorders, paving the way to new therapies. Nonetheless, for the majority of detected genetic variants, the impact on human health is still unknown.

To tackle this problem several computational methods have been developed to predict the effect of single amino acid variants (SAVs) at structural and functional levels (2). One class of algorithms predicts the impact of SAVs on protein stability (3). These approaches, which analyze the protein 3D structure, provide an estimation of the variation of free energy change upon mutation. We recently developed DDGun (4) and ACDC-NN (5), which are among the most accurate methods for predicting the impact of SAVs on protein stability (6). Another class of algorithms is binary classifiers, which discriminate between *pathogenic* and *benign* single nucleotide variants (SNVs) for both coding and non-coding regions. Currently, the state-of-the-art methods achieve a level of accuracy of ~85-90%, and among them, PhD-SNPg (7) was developed the author of this proposal. When focusing on SAVs, all methods described above show better performance when protein structural information is taken into consideration. Mapping mutations onto 3D structures enables atomic-level analyses of 3D-spots in proteins that may be important for stability or engagement in interactions (8). In general, this information may result in a more comprehensive explanation of the mutation impact and of the disease mechanism.

Although the mentioned approaches reach a good level of performance, and are currently incorporated in the variant prioritization pipelines, their application in clinical settings is still not straightforward. At the current stage, the available methods present limitations, mainly due to the complexity of the genotype-phenotype relationship. In general, all the algorithms for predicting the impact of genetic variants are binary classifiers and do not provide information about the type of disease associated with the putative pathogenic variants. Furthermore, many polygenic disorders exist, for which it is well established that gene variants do not cause disease in isolation but rather contribute to perturbing a molecular network, resulting in pathological phenotypes only in combination with other variants (2).

To overcome these limitations, we aim to develop a modular and open computing toolset for the analysis and interpretation of the individual exome that, by integrating different sources of molecular data and prediction algorithms, will estimate the risk of expressing specific classes of RD. This project will rely on the financial and technical support of ELIXIR, a European organization that brings together valuable resources and data centers to create a federated infrastructure for storing and sharing biological data.

The successful completion of our project, integrated with European networks and facilities, is expected to increase our competitiveness in the participation in national and international research projects.

collaboration with the Laboratory of Genetic Medicine of the *"Sant'Orsola"* Hospital (Bologna) which has a solid experience in RD classification and diagnostics.

AIM 3: Implementation of a computational platform for the assessment of RD risk
Using the strategy implemented in the *ContrastRank* algorithm (27), we will define a gene prioritization score based on the occurrence of the putative pathogenic variants in case and control sets. To estimate the background distribution of the putative pathogenic variants across individuals, we will use the genotypes of the ~2,500 individuals from the 1000 Genomes Project. We will use the gene prioritization score to rank the whole exome by considering the subset of genes carrying at least one pathogenic rare variant for each individual. Similar scores can also be calculated on protein clusters with similar functions and/or domains. This procedure results in multiple individual scores based on the different clustering systems. A machine learning method trained on the defined scores will be used to assign a unique probabilistic disease risk to each individual. An alternative analysis will be performed considering for each individual the mutated set of genes in the context of the protein-protein or gene interaction networks. Using an approach similar to that adopted for characterizing the disease networks (28), we will study the segregation of the mutated genes in the same neighborhoods of the human interactome networks.
Finally, all the developed methods and tools will be integrated into reproducible pipelines to be used in secure environment for the analysis, characterization and visualization of genetic variants associated with RDs and the study of their impact on the mutated proteins.
The developed platform will also offer a service for the analysis of whole genome sequencing data: in particular, it will make available to end-users – including clinicians – the computational methods developed for predicting the likelihood of developing a RD based on the occurrence of genetic variants in patients' exomes. Workflows, APIs and containerization of the tools/applications developed in the project will allow download and local usage on different platforms. Our project will include the refactorization of the involved software using a architecture, in which each component is executed in a Docker container on a cluster. All developed code will be open source and available on a GitHub-like platform providing hosting for software projects and version control.
Workflows for the automated analysis of point mutations associated with RDs will be made available on the web platform for download. These will include a series of containers ready to be used by the bioinformatics community. Containers offer a packaging mechanism in which software code and all its dependencies can be abstracted from the environment in which they actually run. This decoupling allows uniform usability of software on any infrastructure, thus better ensuring reproducibility of research results, as well as easy and consistent deployment of applications. Software developed in this project will be packaged up together with its dependencies and examples into Docker or Singularity images. Container-based software will be made available for download and local execution. We will also provide application programming interfaces (APIs) to access the database and query the analysis and prediction methods programmatically. All the tools will be open source; on the other, they will focus on automation, pipeline reproducibility and interaction with FAIR data to enhance sustainability and impact. For these specific tasks we will collaborate with researchers of the local node of the INFN.

It is expected that the application in clinical settings of our tools and platform can result in reduced time and efforts for diagnosis, and in improved treatment strategies, reducing health care costs and increasing patients' benefits.
This project will rely on the financial (CUP: J33C21000210005) and technical (Rare Disease Community) support of ELIXIR, a European organization that brings together valuable resources and data centers to create a federated infrastructure for storing and sharing biological data.
Finally, the successful completion of our project, integrated with European networks and facilities, is expected to increase our competitiveness in the participation in national and international research projects.

References

1. S. Nguengang Wakap *et al.*, *Eur J Hum Genet*. 28, 165–173 (2020).
2. E. Capriotti, K. Ozturk, H. Carter, *Wiley Interdiscip Rev Syst Biol Med*. 11, e1443 (2019).
3. T. Sanavia *et al.*, *Comput Struct Biotechnol J*. 18, 1968–1979 (2020).
4. L. *Montanucci et al., Nucleic Acids Res.* DOI:10.1093/nar/gkac325 (2022).
5. C. Pancotti *et al.*, *Genes (Basel)*. 12, 911 (2021).
6. C. Pancotti *et al.*, *Brief Bioinform*. 23, bbab555 (2022).
7. E. Capriotti, P. Fariselli, *Nucleic Acids Res*. 45, W247–W252 (2017).
8. E. Capriotti, R. B. Altman, *BMC Bioinformatics*. S4, S3 (2011).
9. M. J. Landrum et al., *Nucleic Acids Res.* 48, D835–D844 (2020).
10. UniProt Consortium, *Nucleic Acids Res.* 47, D506–D515 (2019).
11. S. Gore et al., *Structure.* 25, 1916–1927 (2017).
12. J. Jumper et al., *Nature*. 596, 583–589 (2021).
13. J. Mistry et al., *Nucleic Acids Res*. 49, D412–D419 (2021).
14. M. Ashburner et al., *Nat Genet.* 25, 25–9 (2000).
15. P. Porras et al., Nat Commun. 11, 6144 (2020).
16. D. Szklarczyk et al., *Nucleic Acids Res*. 47, D607–D613 (2019).
17. M. A. Freeberg et al., *Nucleic Acids Res*. 50, D980–D987 (2022).
18. G. A. Van der Auwera et al., *Curr Protoc Bioinformatics*, in press, doi:10.1002/0471250953.bi1110s43.
19. D. C. Koboldt, *Genome Medicine*. 12, 91 (2020).
20. R. Poplin et al., *Nat Biotechnol*. 36, 983–987 (2018).
21. H. Yang, K. Wang, *Nat Protoc*. 10, 1556–1566 (2015).
22. K. J. Karczewski et al., *Nature*. 581, 434–443 (2020).
23. T. Wang et al., *Nat Commun*. 11, 4932 (2020).
24. A. R. Cardoso et al., *Human Genomics*. 13, 31 (2019).
25. C. S. Leblond et al., *Mol Cell Neurosci*. 113, 103623 (2021).
26. C. Chen et al., *Neurosci Lett*. 685, 96–101 (2018).
27. R. Tian, M. K. Basu, E. Capriotti, *Bioinformatics*. 30, i572-578 (2014).
28. J. Menche et al., *Science*. 347, 1257601 (2015).

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI
FARMACIA E BIOTECNOLOGIE

Oggetto: Piano formativo assegno BIR 2022 cofinanziato fondi Emidio Capriotti (CUP J33C21000210005)

Per ogni obiettivo del progetto di ricerca, Il piano formativo relativo all'assegno BIR 2022 cofinanziato al 50% con i fondi EMBL-ELIXIR prevede le seguenti le attività formative:

1. **Raccolta dati per la valutazione dei metodi per l'analisi del genoma umano**
   Nella prima fase del progetto (mesi 1-4), l'assegnista acquisirà la capacita di interrogare automaticamente database disponibili online tramite l'utilizzo di programmi. In particolare verranno estratte informazioni dalla banca dati ClinVar e integrati con quelle strutturali e funzionali disponibili sulle banche dati PDB e UniProt. Questi dati verranno combinati con le annotazioni dei genotipi delle diverse malattie rare di origine genetica riportate nella banca data Orphanet. Per costruire una rete di geni associati ad ogni classe di malattia verranno considerati dati di interazioni proteina-proteina. Il risultato di questa prima fase dell'attività consiste nello sviluppo di una collezione di mutazioni di singoli amino acidi e relativi geni associati a malattie rare e le reti geniche coinvolte.

2. **Valutazione e ottimizzazione dei metodi per la predizione i varianti genetiche patogene associate a malattie rare**
   Dal mese 5 al mese 8, l'assegnista istallerà e utilizzerà su infrastrutture computazionali, tramite accesso remoto, i metodi classici per la predizione dell'impatto delle mutazioni amino ecidiche sulla stabilità (FOLDX, DDGun) e sulla funzione (CADD, FATHMM, PhD-SNPg) delle proteine. Una volta ottenute le predizioni, dal mese 9 al mese 12, l'assegnista analizzerà le predizioni per identificare le soglie che permettono di ottimizzare le predizioni dei diversi metodi su specifiche classi di malattia. In particolare il lavoro si focalizzerà su dati di sequenziamento relativi a malattie del neurosviluppo depositati nella banca dati EGA. Questi dati verranno confrontati con quelli messi a disposizione dal consorzio *1000 Genomes*.

3. **Sviluppo di metodi e piattaforme computazionale per la predizione del rischio di sviluppare malattie rare**
   Per raggiungere l'obbiettivo finale del progetto, dal mese 13 al 18, l'assegnista utilizzerà i dati analizzati durante il primo anno del progetto per sviluppare metodi probabilistici capaci di valutare per ogni individuo il rischio di sviluppare una malattia del neurosviluppo a partire dalle mutazioni identificate nell'esoma. Nella fase finale del progetto, dal mese 19 al 24, l'assegnista svilupperà una piattaforma computazionale che permetta l'analisi automatica dell'esoma umano. Tale piattaforma sarà resa disponibile in modalità "*open access*" e testata anche in ambienti computazionali protetti dalla normativa sulla privacy.

   Nell'ambito del progetto l'assegnista avrà la possibilità di collaborare con ricercatori del Laboratorio di Genetica Medica dell'ospedale Sant'Orsola di Bologna e con ricercatori del nodo locale dell'INFN per raggiungere gli obiettivi 2 e 3 del progetto.

Emidio Capriotti
Via Selmi 3 | 40126 Bologna | Italia | Tel. + 39 051 2094303 | emidio.capriotti@unibo.it

Emidio Capriotti - Selected articles

BIR 2022

1. *Montanucci L†,* ***Capriotti E†,*** *Birolo G, Benevenuta S, Pancotti C, Lal D, Fariselli P** (2022). DDGun: an untrained predictor of protein stability changes upon amino acid variants. **Nucleic Acids Research**. DOI:10.1093/nar/gkac325.

2. ***Capriotti E*,*** *Fariselli P** (2022). Evaluating the relevance of sequence conservation in the prediction of pathogenic missense variants. **Human Genetics**. DOI:10.1007/s00439-021-02419-4.

3. *Petrosino M, Novak L, Pasquo A, Chiaraluce R, Turina P,* ***Capriotti E*,*** *Consalvi V*.* (2021). Analysis and interpretation of the impact of missense variants in cancer**. International Journal of Molecular Sciences**. 22:5416. DOI:10.3390/ijms22115416.

4. *Benevenuta S, Capriotti E*, Fariselli P**. (2020). Calibrating variant-scoring methods for clinical decision making. **Bioinformatics**. 36:5709-5711.

5. *Savojardo C†, Petrosino M†, Babbi G, Bovo S, Corbi-Verge C, Casadio R, Piero Fariselli P, Folkman L, Garg A, Karimi M, Katsonis P, Kim PM, Lichtarge O, Martelli PL, Pasquo A, Pal D, Shen Y, Strokach AV, Turina P, Zhou Y, Andreoletti G, Brenner S, Chiaraluce R, Consalvi V,* ***Capriotti E*.*** (2019). Evaluating the predictions of the protein stability change upon single amino acid substitutions for the FXN CAGI5 challenge. **Human Mutation**. 40: 1392-1399.

6. ***Capriotti E*,*** *Fariselli P*.* (2017). PhD-SNP^{g}: A webserver and lightweight tool for scoring single nucleotide variants. **Nucleic Acids Research.** 45: W247-W252.

7. *Tian R, Basu MK,* ***Capriotti E*.*** (2014). ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. **Bioinformatics.** 30: i572-i578.

8. ***Capriotti E*,*** *Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R** (2013). WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. **BMC Genomics.** 14 Suppl 3:S6.