Title of your report

ParBigMen: ParSMURF application to Big genomic and epigenomic data for the detection of pathogenic variants in Mendelian diseases

Research institution

[1] AnacletoLab – Dipartimento di Informatica, Università degli Studi di Milano (Italy)

Principal Investigator

Alessandro Petrini[1], Giorgio Valentini[1]

Researchers

Tiziana Castrignanò[2], Peter Robinson[3], Marco Frasca[1], Elena Casiraghi[1], Sara Bonfitto[1]

Project partners

[2] Department of Ecological and Biological Sciences (DEB) - Università della Tuscia, Viterbo (Italy)

[3] The Jackson Laboratory for Genomic Medicine, Farmington - CT (USA)

SuperMUC-NG project ID(s) of the projects you report in this article

pn29lu

Introduction

Variant identification and analysis of Next-Generation Sequencing (NGS) data play a central role in Genomic and Personalized Medicine. In this context, disease-associated variants occurring in protein-coding areas of the DNA are well studied, but the understanding of the impact of variants occurring in the non-coding regions of the genome is for most part incomplete. However, the scientific community shifted its attention towards the understanding of the latter regions, as several recent studies state that most of the potential pathogenic and deleterious variants do not lie in the coding areas of the genome.

We contributed to this emerging field of Genomic Medicine by developing machine learning tools for the detection of pathogenic and deleterious variants in the non-coding genome. This task has been proven to be particularly challenging for several reasons, most of which are related with the sparsity of pathogenic mutations which are outnumbered by neutral variants: when tackled with an automated learning approach, this ultimately translates to a high unbalance between classes of examples to be learned, leading to a very challenging classification problem.

With HyperSMURF [1] and ParSMURF [2] we proposed two state-of-the-art Machine Learning solutions for dealing with such ill-posed datasets; in particular, HyperSMURF is the evolution of ReMM (Regulatory Mendelian Mutation), the Machine Learning core of Genomiser [3] used for the diagnosis and discovery of genetic variants causative of Mendelian disorders.

The main objective of *ParBigMen: ParSMURF application to Big genomic and epigenomic data for the detection of pathogenic variants in Mendelian diseases* is to improve the results obtained so far for providing the scientific community with ever more reliable means for the prioritization of variants associated with Mendelian diseases.

The core idea of the project is to improve the prediction performance of the base classifier by using a largely expanded dataset. This improvement of data coverage is done by i) adding new epigenomic features with data from publicly available repositories for increasing the characterization of each sample, and ii) adding newly discovered deleterious and pathogenic samples.

However, this increase of data cardinality, along with the high complexity of ParSMURF and the fine tuning of its learning parameters, comes with an increase of computational costs, making this project feasible only on a Tier-0 supercomputing facility.

To summarize, the goals of *ParBigMen* are:

1) the release of a new highly parallel version of ParSMURF (called ParSMURF-NG) able to scale with big data and to fully exploit the High-Performance Computing architectures for relevant prediction problems in the context of Personalized and Precision Medicine.

2) the application ParSMURF-NG to big omics data, where many features will be investigated and used to predict pathogenic variants.

3) the evaluation and release of new ReMM scores to the entire scientific community for the prioritization of pathogenic and deleterious SNVs.

Results and Methods

The main tool developed for this project is ParSMURF-NG. It was specifically developed to improve the scalability and computational efficiency of ParSMURF, thus meeting the computational demand required by *ParBigMen*.

Its scalability and efficiency have been evaluated on three supercomputing systems (Marconi KNL and SKL of Cineca, HLRS Apollo HAWK of HPC Center Stuttgart and SuperMUC-NG) reaching >70% of efficiency even with 128 computing nodes (up to 32k computing cores on Marconi KNL). Fig. 1 reports the strong scaling speed-up and efficiency of ParSMURF-NG on a benchmark dataset of 30M samples and 50 features, also comparing its performance on the three systems. Although showing a slightly lower speed-up compared to Marconi and Hawk, SuperMUC-NG excelled in the overall execution time, being twice faster than Marconi and approximately four times than Hawk.

ParSMURF-NG is entirely programmed in C++ and is highly optimized for CPUs featuring a high core count such as Intel Xeon Phi processors. All inter-process communication is managed through calls to the MPI library; access to the filesystem is managed in the same way to properly exploit both the shared file system and the high-speed intercommunication infrastructure of contemporary HPC systems. As such, we measured that a pool of 128 ParSMURF-NG processes distributed on 128 nodes of SuperMUC-NG can read the input dataset at a rate of 500 MB/s each.

ParSMURF-NG is distributed as source code and is publicly available at [4].

To predict the pathogenicity of genomic variants we trained the classifier of ParSMURF-NG with a dataset consisting of 14 million variants characterized with 26 heterogeneous features. Of all these variants, only 406 are associated with known Mendelian diseases. This dataset has been used in [3] for evaluating the ReMM scores and in *ParBigMen* represents the baseline for comparison.

To provide a more informative dataset to the classifier of ParSMURF-NG we expanded this dataset by adding 80 new pathogenic variants and by characterizing each variant with new epigenomic features extracted from the International Human Epigenomic Consortium data portal. We created more than 500 new features, hence greatly increasing the size of the dataset.

We applied several feature selection strategies to narrow down the number of significant features for this problem: in particular, we considered several classic indexes used for evaluating correlation (Spearman and Pearson correlation coefficient), independence of distribution (Mann-Whitney and Kruskal-Wallis) as well as multivariate methods (MRMR) and a wrapper method (ParSMURF-NG feature importance).

After selecting the set of most significant features, we used SuperMUC-NG for the highly computing intensive operation of finding the set of best learning hyper-parameters of ParSMURF-NG. This model selection task is crucial to properly train a classifier so that it can deliver reliable predictions.

Specifically, in *ParBigMen*, for each combination of the selected features we exhaustively explored a search space of 1440 configurations. ParSMURF-NG distributed the computation across 2560 nodes of SuperMUC-NG, delivering the set of best hyper-parameters in less than 12 hours. We estimated that each model selection task performed on a single machine would have taken almost two years to deliver the same results.

Ongoing Research / Outlook

Thanks to the computational power of the SuperMUC-NG system, we managed to find the minimal set of genomic features and the set of best learning hyper-parameters in a feasible amount of time. As briefly reported in Table 1, we increased the performance of the classifier in predicting pathogenic and deleterious variants, by selecting a novel set of epigenomic and conservation features and by finely tuning the hyper-parameteres of our highly parallelized algorithm.

Given the promising results obtained in *ParBigMen*, we plan to expand the experimental set-up by including new highly parallelized methods for multi-variate feature selection and by possibly exploring novel variant features to achieve novel insights into candidate pathogenic variants associated with Mendelian diseases.

References and Links

[1] Schubach M., Re M., Robinson P.N., Valentini G., “Imbalance Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants”. Scientific Reports 2017;7(1):2959.

[2] Petrini A., Mesiti M., Schubach M., Frasca M., Danis D., Re M., Grossi G., Cappelletti L., Castrignanò T., Robinson P. N., Valentini G., “parSMURF, a high-performance computing tool for the genome-wide detection of pathogenic variants”, GigaScience, vol. 9, 05 2020.

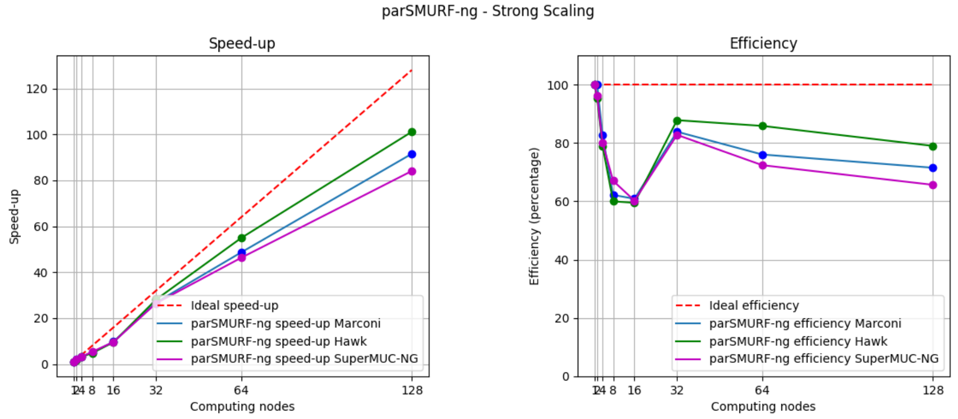
[3] Smedley D., Schubach M., Jacobsen J.O.B., Köhler S., Zemojtel T., Spielmann M., et al. “A Whole Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease”. The American Journal of Human Genetics 2016 sep;99(3):595–606

[4] ParSMURF-NG on AnacletoLab GitHub page: https://github.com/AnacletoLAB/parSMURF-NG

Table 1: comparison of the classifier performance in AUROC and AUPRC when trained with the original and the improved dataset using ParSMURF-NG

	AUROC	AUPRC
Baseline dataset	0.99361	0.34270
New dataset	0.99393	0.42410

Figure 1: speed up and efficiency of ParSMURF-NG on Marconi KNL, Apollo HAWK and SuperMUC-NG supercomputing systems, when measured on a strong scaling setup-up over a synthetic dataset


parSMURF-ng - Strong Scaling
Speed-up
Speed-up
Computing nodes
Ideal speed-up
parSMURF-ng speed-up Marconi
parSMURF-ng speed-up Hawk
parSMURF-ng speed-up SuperMUC-NG
Efficiency
Efficiency (percentage)
Computing nodes
Ideal efficiency
parSMURF-ng efficiency Marconi
parSMURF-ng efficiency Hawk
parSMURF-ng efficiency SuperMUC-NG