ULB FACULTÉ DES SCIENCES
UNIVERSITÉ LIBRE DE BRUXELLES

VUB VRIJE UNIVERSITEIT BRUSSEL

Towards multivariant pathogenicity predictions

Using machine-learning to directly predict and explore disease-causing oligogenic variant combinations

Thesis submitted by Sofia PAPADIMITRIOU

in fulfilment of the requirements of the PhD Degree in Sciences (ULB - “Docteur en Sciences”) and in Sciences (VUB)
Academic year 2019-2020

Supervisors: Professor Tom LENAERTS (Université Libre de Bruxelles)

Machine Learning Group

and Professor Ann NOWÉ (Vrije Universiteit Brussel)

Artificial Intelligence Lab

Thesis jury:

Gianluca BONTEMPI (Université Libre de Bruxelles, Chair)
Guillaume SMITS (Université Libre de Bruxelles, Secretary)
Coen DE ROOVER (Vrije Universiteit Brussel)
Sonia VAN DOOREN (Vrije Universiteit Brussel)
Elfride DE BAERE (Ghent University)
Christian GILISSEN (Radboud UMC)

fnrs LA LIBERTÉ DE CHERCHER
icity.brussels THE ICT HEART OF EUROPE
innoviris.brussels empowering research
RÉGION DE BRUXELLES-CAPITALE BRUSSELS HOOFDSTEDELIJK GEWEST
Union Européenne Europese Unie

This PhD thesis has been conducted and written under the supervision of prof. dr. Tom Lenaerts (Université Libre de Bruxelles) and prof. dr. Ann Nowé (Vrije Universiteit Brussel).

The public defense took place on 15th September 2020, at Université Libre de Bruxelles.

Thesis abstract

English

The emergence of statistical and predictive methods able to analyse genomic data has revolutionised the field of medical genetics, allowing the identification of disease-causing gene variants (i.e. mutations) for several human genetic diseases. Although these approaches have greatly improved our understanding of Mendelian «one gene – one phenotype» genetic models, studying diseases related to more intricate models that involve causative variants in several genes (i.e. oligogenic diseases) still remains a challenge, either due to the lack of sufficient methodologies and disease-specific cohorts to study or the phenotypic complexity associated with such diseases. This situation makes it difficult to not only understand the genetic mechanisms of the disease, but to also offer proper counseling and support to the patient. Until recently, no specialized predictive methods existed to directly predict causative variant combinations for oligogenic diseases. However, with the advent of data on variant combinations in gene pairs (i.e. bilocus variant combinations) leading to disease, collected at the Digenic Diseases Database (DIDA), we hypothesized that the transition from single to variant combination pathogenicity predictors is now possible.

To investigate this hypothesis, we organised our research on two main routes. At first, we developed an interpretable variant combination pathogenicity predictor, called VarCoPP, for gene pairs. For this goal, we trained multiple Random Forest algorithms on pathogenic bilocus variant combinations from DIDA against neutral data from the 1000 Genomes Project and investigated the contribution of the incorporated variant, gene and gene pair features to the prediction outcome. In the second part, we explored the usefulness of different gene pair burden scores based on this novel predictive method, in discovering oligogenic signatures in neurodevelopmental diseases, which involve a spectrum of monogenic to polygenic cases. We performed a preliminary analysis on the Deciphering Developmental Diseases (DDD) project containing exome data of 4195 families and assessed the capability of our scores in supporting already diagnosed monogenic cases,

discovering significant pairs compared to control cases and linking patients in communities based on the genetic burden they share, using the Leiden community detection algorithm.

The performance of VarCoPP shows that it is possible to predict disease-causing bilocus variant combinations with good accuracy both during cross-validation and when testing on new cases. We also show its relevance for disease-related gene panels, and enhanced its clinical applicability by defining confidence zones that guarantee with 95% or 99% probability that a prediction is indeed a true positive, guiding clinical researchers towards the most relevant results. This method and additional biological annotations are incorporated in an online platform called ORVAL that allows the prediction and exploration of candidate disease-causing oligogenic variant combinations with predicted gene networks, based on patient variant data. Our preliminary analysis on the DDD cohort shows that - although all bi-locus burden scores show advantages, disadvantages and certain types of biases - taking the maximum pathogenicity score present inside a gene pair seems to provide, at the moment, the most unbiased results. We also show that our predictive methods enable us to detect patient communities inside DDD, based exclusively on the shared pathogenic bi-locus burden between patients, with more than half of these communities containing enriched phenotypic and molecular pathway information. Our predictive method is also able to bring to the surface genes not officially known to be involved in disease, but nevertheless, with a biological relevance, as well as a few examples of potential oligogenicity inside the network, paving the way for further exploration of oligogenic signatures for neurodevelopmental diseases.

Français

L'émergence de méthodes statistiques et prédictives capables d'analyser les données génomiques a révolutionné le domaine de la génétique médicale, permettant l'identification de variants de gènes pathogènes pour plusieurs maladies génétiques humaines. Bien que ces méthodes aient considérablement approfondi notre compréhension des modèles mendéliens "un gène - un phénotype", l'étude des maladies liées à des modèles plus complexes impliquant des variants de plusieurs gènes (c'est-à-dire les maladies oligogéniques) reste un défi, notamment en raison soit d'un manque de cohortes suffisamment grandes et dont les patients souffriraient tous de la même maladie soit de l'émergence de cas présentant une pénétrance incomplète et un manque de variabilité phénotypique chez les patients. Cette situation entrave non seulement la compréhension des mécanismes génétiques de la maladie, mais aussi de la capacité à offrir un conseil et un soutien appropriés au patient. Il n'existait auparavant pas de méthodes prédictives spécialisées permettant de déterminer directement lescombinaisons de variants responsables de maladies oligogéniques. Cependant, avec la récentedisponibilité de données sur les combinaisons de variants dans des paires de gènes (c'est-à-dire les combinaisons de variants bilocales) conduisant à une maladie génétique, recueillies dans la base de données sur les maladies digénétiques (Digenic Diseases Database, DIDA), nous avons émis l'hypothèse que la transition depuis des prédicteurs de pathogénicité à un seule variant vers des combinaisons de variants est maintenant à notre portée.

Afin d'étudier cette hypothèse, nous avons organisé notre recherche en deux axes principaux. Dans un premier temps, nous avons développé un prédicteur interprétable de la pathogénicité d'une combinaison de variants, appelé VarCoPP, pour des paires de gènes. Dans ce but, nous avons entraîné plusieurs modèles d'algorithmes de forêts aléatoires sur des combinaisons de variantsbilocales pathogènes provenant de DIDA groupées avec des combinaisons neutres du projet 1000 Génomes et nous avons étudié la contribution des caractéristiques des variants, des gènes et des paires de gènes en rapport avec le résultat de la prédiction. Pour la deuxième partie, nous avons exploré l'utilité de différents indices de poids de paires de gènes basés sur cette nouvelle méthode de prédiction, pour découvrir les signatures oligogéniques dans les maladies neurodéveloppementales (MND) qui démontrent un éventail de cas monogéniques et polygéniques. Nous avons effectué une analyse préliminaire sur le projet "Deciphering Developmental Diseases" (DDD) comprenant des données sur les exomes de 4195 familles etavons évalué la capacité de nos scores à corroborer des cas monogéniques déjà identifiés, à découvrir des paires significatives par rapport aux cas de référence et à associer des patients dans des communautés en fonction de la charge génétique qu'ils partagent, grâce à l'algorithme de détection de communautés

de Leiden.

Les performances de VarCoPP montrent qu'il est possible de prévoir avec une précision adéquate les combinaisons de variants bilocales pathogènes, tant lors de la validation croisée que lors des tests sur les nouveaux cas. Nous avons également montré sa pertinence pour les panels de gènes liés à une maladie spécifique et amélioré son applicabilité clinique en définissant des zones de confiance qui garantissent avec une probabilité de 95 ou 99% qu'une prédiction est effectivement positive, aidant les chercheurs cliniciens à identifier les résultats les plus pertinents. Cette méthode, ainsi que des annotations biologiques supplémentaires, sont intégrées dans une plateforme en ligne nommée ORVAL qui permet de prédire et d'explorer les combinaisons oligogéniques de variants causant potentiellement des maladies grâce à des réseaux de gènes, sur la base des données des variants génétiques des patients. Notre analyse préliminaire sur la cohorte DDD montre que - bien que tous les scores de charge bi-locus présentent des avantages, des inconvénients et certains types de biais - le fait de prendre le score maximal de pathogénicité présent à l'intérieur d'une paire de gènes semble fournir, pour le moment, les résultats les plus biaisés. Nous montrons également que nos méthodes prédictives nous permettent de détecter des communautés de patients au sein de DDD, en se basant exclusivement sur la charge de bi-locus pathogène partagée entre les patients, avec plus de la moitié de ces communautés contenant des informations phénotypiques et moléculaires enrichies. Notre méthode prédictive est également capable de faire remonter à la surface des gènes non officiellement connus pour être impliqués dans des maladies, mais néanmoins avec une pertinence biologique, ainsi que quelques exemples d'oligogénicité potentielle à l'intérieur du réseau, ouvrant la voie à une exploration plus approfondie des signatures oligogéniques pour les maladies neurodéveloppementales.

Nederlands

Vooruitgang in de ontwikkeling van statistische en voorspellende technieken voor de identificatie van ziekteveroorzakende genvarianten voor verschillende genetische aandoeningen op basis van genetische data, heeft het onderzoeksveld van de medische genetica gerevolutioneerd. Hoewel deze technieken sterk hebben bijgedragen tot het begrijpen van het Mendeliaanse "één gen – één fenotype" genetische model, vormt de studie van ziekten gerelateerd aan complexere overervingsmodellen, waarbij causale varianten in verschillende genen een rol spelen, nog steeds een uitdaging. Deze uitdaging volgt aan de ene kant uit een gebrek aan voldoende ziekte specifieke behandelingsgroepen om te bestuderen en aan de andere kant door de aanwezigheid van gevallen met onvoldoende penetrantie in de samenleving als ook een grote fenotypische variabiliteit onder de patiënten. Deze situatie maakt het niet alleen moeilijk om de genetische mechanismen van een ziekte te begrijpen, maar ook om een goede begeleiding en ondersteuning te voorzien voor de patiënten. Tot voor kort bestonden er geen gespecialiseerde methoden om causatieve varianten voor oligogene ziekten rechtstreeks te voorspellen. Dankzij nieuwe gegevens over variantencombinaties in genenparen (d.w.z. bilocus variantencombinaties) die tot een ziekte leiden, verzameld in de Digenic Diseases Database (DIDA), hebben we de hypothese gesteld dat de overgang van enkelvoudige- naar meervoudige pathogeniteitsvoorspellers nu wel mogelijk is.

Het doctoraatsonderzoek dat hier wordt voorgelegd levert twee bijdragen die deze hypothese bevestigen. In eerste plaats hebben we de eerste, interpreteerbare variantencombinatie pathogeniteitsvoorspeller voor genenparen ontwikkeld, genaamd VarCoPP. Hiervoor hebben we meerdere Random Forest-algoritmen getraind op basis van pathogene bilocus-variantencombinaties beschikbaar in DIDA en neutrale data van het 1000 Genomen Project. We hebben de invloed op de voorspelling van de ingebouwde variant-, gen- en gen paar-karakteristieken onderzocht alsook de algemene kwaliteit en de klinische relevantie van de voorspellingen. In het tweede deel hebben we het nut onderzocht van verschillende belastingsscores op basis van deze nieuwe voorspellende methode voor het ontdekken van oligogene vingerafdrukken bij neurologische ontwikkelingsziekten, waarbij het volledige spectrum van monogene tot polygene situaties mogelijk zijn. We voerden deze eerste analyse uit op de genetische data beschikbaar via het Deciphering Developmental Diseases (DDD)-project. Via dat project konden we beschikken over de exomen van 4195 families en beoordeelden we het vermogen van onze scores om reeds gediagnosticeerde monogene gevallen te bevestigen, nieuwe significante paren te ontdekken in vergelijking met controlegevallen en patiënten te verdelen over groepen op basis van de genetische belasting die ze delen.

De prestaties van VarCoPP tonen aan dat het mogelijk is om ziekteverwekkende combinaties van bilocus-varianten met een goede nauwkeurigheid te voorspellen, zowel tijdens de validatie als bij het testen van nieuwe gevallen. We tonen ook de relevantie voor ziektegerelateerde genengroepen en verbeterden de klinische toepasbaarheid ervan door betrouwbaarheidsintervallen te definiëren die met een waarschijnlijkheid van 95% of 99% garanderen dat een voorspelling een waarachtig positief resultaat is. Dit laatste is essentieel omdat het klinische onderzoekers toelaat om naar de meest relevante resultaten te kijken. Deze methode en aanvullende biologische annotaties zijn opgenomen in een online platform genaamd ORVAL dat – gebaseerd op patiëntvariantgegevens – de voorspelling en verkenning mogelijk maakt van kandidaat-ziekteveroorzakende oligogene variantencombinaties op basis van verschillende, gelaagde visualisaties. Onze analyse van de DDD-patiëntengroep laat zien dat ondanks alle voor- en nadelen van de verschillende aggregatiescores, het nemen van de maximale aggregatiescore in een genenpaar de beste resultaten blijkt te geven. We laten ook zien dat deze aggregatiescore ons in staat steelt om patiëntengroepen in DDD te detecteren, uitsluitend gebaseerd op de gedeelde pathogene aggregatiescore tussen patiënten, waarbij meer dan de helft van deze gemeenschappen verrijkte fenotypische en moleculaire paden informatie bevatten. Onze methodes identificeren daarbij ook nieuwe genen waarvan niet officieel erkend is dat ze bij een ziekte betrokken zijn, maar niettemin een biologische relevantie hebben. Op deze manier ontdekten we voorbeelden van mogelijke oligogeniciteit binnen het netwerk, wat dus de basis legt voor verder onderzoek naar oligogene eigenschappen van neurologische ontwikkelingsstoornissen.

Acknowledgements

This thesis concludes four years of an amazing journey, both on a personal and professional level. For many, life when doing a PhD can be remembered as a period with a lot of anxiety, uncertainty and feelings of self-doubt. For sure, during my PhD things were not often straightforward and sometimes I felt frustration and stress, especially having to finish my thesis during a global pandemic. Nevertheless, I can only consider myself as blessed that I can recall it as a very positive experience. Doing my PhD involved moving to a new place, Brussels, that I can now call my home, working on a project that I was really motivated and interested in, and meeting amazing people from all over the world. There are so many people to whom I owe gratitude for reaching this far and a big *thank you*.

To my two supervisors, prof. Tom Lenaerts from the ULB side and prof. Ann Nowé, from the VUB side. I would like to thank Tom for for his support, encouragement and for his kind words every time I made mistakes, and for arranging our online morning lab coffee sessions during the uncertain times of a pandemic, when we were all confined at home, so that everyone could feel that they can have a person to talk to. I would like to thank Ann for her valuable comments on my work and her continuous support and her quick help in everything I needed.

Of course, to the members of my jury for accepting to evaluate my work. To prof. Gianluca Bontempi, who was also the president of my commité d'accompagnement, for his valuable comments, as well as his advice during the development of our bi-locus burden scores. To prof. Coen De Roover, prof. Elfride de Baere and prof. Christian Gilissen, for accepting to read my work with a critical mind. Last but not least, to prof. Guillaume Smits and prof. Sonia van Dooren whose feedback, being clinical researchers, was very valuable during the development of our predictive methods. I would also like to thank prof. Smits for his continuous interest and input in my work, and for his spot-on questions during my last rehearsal, one day before taking my interview for the FNRS-FRIA scholarship, which contributed to me obtaining it.

To prof. Wim Wranken, my study advisor at VUB, for his valuable input on my work and his questions that made me re-think my strategy on the bi-locus burden scores. To

prof. Patrick Mardulyn, member of my commité d'accompagnement, for his very kind input and positive feedback during my yearly evaluations.

To FNRS for granting me the 3-year FRIA scholarship, to Innoviris and icity.brussels for providing our team with the opportunity to work on such an interesting project, and to VUB for financing a whole year of my PhD.

To our oligogenic team at IB^2, as I believe that someone cannot have better teammates. To my daily office-mate Alexandre Renaux, who made my stay in the office joyful, productive and full of music, as well as for his excellent advancements on ORVAL that made our work more visible to the scientific community. To Charlotte Nachtegael for our continuous counter-support at a personal level, as well as a professional level with our common struggles on the DDD cohort. To Nassim Versbraegen for his valuable help in creating our crucial pipelines that enabled us to proceed our work and establish important collaborations. To Simon Boutry who helped us with our ORVAL project and to Arnau Dillen for his excellent work on moving DIDA towards OLIDA.

To Andrea Gazzo, who was my mentor when I first arrived at IB^2 during my internship five years ago, and who warmly supported me during my first struggles with machine learning. To Jelena Grujic for her very valuable input on my community detection methodology, a topic that I recently started to gain more experience on.

To all my colleagues at IB^2 and the Machine Learning Group for the excellent professional environment that they create, which made me want to return to the office every day. To Nathaniel Monpère and Stefan Vet for our PhD support group. To Elias Fernandez for his friendship and counter-support, as for some reason we always ended up writing scholarship proposals, and in the end our theses, at the same time.

To my dear friends Marianna and Elsa with whom I have been sharing so many great memories since the beginning of my master in the Netherlands until now, always supporting each other during our own PhDs. To my childhood friends from Greece. The constant contact with them, even being far away, made my days happier and gave me strength to continue.

The most special part of my gratitude, to my parents and family for their endless, selfless support throughout all these years and for helping me emotionally and financially, leaving many of their own personal needs behind, in order for me to be where I am today. To my dear aunt and uncle in Brussels who were always eager to help me like their own child since the beginning of my stay in Brussels. To my grandparents for their very kind soothing words throughout all this period and their eagerness to learn to use Skype by themselves while I have been away, so that they can reach and talk to me.

Contents