Instituto_Nacional de Saúde Doutor Ricardo Jorge
Observações_ Boletim Epidemiológico
2021 número 30
www.insa.pt
artigos breves_ n. 10
_Medicina personalizada

_Aplicação de métodos de aprendizagem automática em grafos de conhecimento para medicina personalizada

Application of machine learning methods for knowledge graphs to personalized medicine

Joana Vilela[1,2], Muhammad Asif[1,2], Ana Rita Marques[1,2], João Xavier Santos[1,2], Célia Rasga[1,2], Astrid Vicente[1,2], Hugo Martiniano[1,2]

hugo.martiniano@insa.min-saude.pt

(1) Departamento de Promoção da Saúde e Prevenção de Doenças Não Transmissíveis, Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal

(2) Biosystems and Integrative Sciences Institute. Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

_Resumo

A medicina personalizada é um modelo de prática médica que utiliza o perfil fenotípico e genotípico do indivíduo para melhorar a precisão do diagnóstico, a eficácia terapêutica ou a prevenção de doenças. Neste sentido, a enorme quantidade de dados gerados ao longo dos últimos anos na área biomédica tem contribuído para uma melhor compreensão dos determinantes genéticos de várias patologias e, consequentemente, para a implementação de práticas de medicina personalizada em várias áreas, por exemplo na Oncologia e no âmbito das doenças raras. No entanto, ainda subsistem desafios significativos, nomeadamente no que diz respeito à integração de dados biomédicos oriundos de fontes heterogéneas e na obtenção de informação clinicamente relevante. Este trabalho descreve uma abordagem que usa métodos de aprendizagem automática aplicados a um Grafo de Conhecimento (GC) biomédico como um meio para integrar informação armazenada em bases de dados diversas. Este GC contém relações entre genes, doenças e outras entidades biológicas, extraídas de três bases de dados: Ensembl, DisGeNET e Gene Ontology. Neste trabalho exploramos o potencial dos métodos de aprendizagem automática em grafos para produzir informação clinicamente relevante e descrevemos a aplicação desta metodologia à previsão de associações gene-doença. Mostramos ainda que as principais associações gene-doença previstas por esta abordagem podem ser confirmadas em bases de dados externas ou já foram previamente identificadas na literatura.

_Abstract

Personalized medicine is a model of medical practice that uses an individual's phenotypic and genotypic profile to improve diagnostic accuracy, therapeutic efficacy or disease prevention. The huge amount of data generated over the last few years in the biomedical area has contributed to a better understanding of the genetic determinants of various pathologies and, consequently, to the implementation of Personalized Medicine practices in various areas, for example in Oncology and in the field of rare diseases. However, significant challenges remain, namely with regard to the integration of biomedical data from heterogeneous sources to obtain clinically relevant information. This work describes an approach that uses machine learning methods applied to a biomedical Knowledge Graph (KG) as a means to integrate information stored in different databases. To build the KG, three databases were used: Ensembl, DisGeNET and Gene Ontology. This KG contains relationships between genes, diseases, and other biological entities. In this work, we explore the potential of automatic graph learning methods to produce clinically relevant information and describe the application of this methodology to the prediction of gene-disease associations. We show that the main gene-disease associations predicted by this approach can be confirmed in external databases or have been previously identified in the literature.

_Introdução

A medicina personalizada é uma prática médica emergente baseada na observação de que o contexto individual de doença é único, pois é moldado pelo perfil molecular e fisiológico do indivíduo, em articulação com fatores comportamentais e exposição ambiental [1,2]. Compreender os determinantes genéticos de doenças é um passo essencial para a aplicação de abordagens de medicina personalizada. Neste trabalho, exploramos o uso de métodos de aprendizagem automática em grafos de conhecimento (GC) [3,5] como uma ferramenta para modelar as relações entre entidades biológicas, tais como genes e doenças, e obter informações sobre associações gene-doença que possam ser úteis em medicina personalizada. Um grafo G (V, E), é uma estrutura matemática constituída por um conjunto de vértices (V) e um conjunto de arestas (E). Muitos conjuntos de dados são naturalmente representados por grafos, nomeadamente as redes, ou, mais genericamente, qualquer conjunto de entidades com ligações entre elas. Um GC é um multigrafo heterogéneo dirigido, onde cada vértice ($v \in V$) representa uma entidade e cada aresta ($e \in E$) uma relação. As entidades e as relações num GC são organizadas em conjuntos de tripletos (h, r, t), onde h é a entidade principal, r é a relação e t é

a entidade final. Cada tripleto pode ser interpretado como a representação de um facto, no qual a entidade principal (ou sujeito) está relacionada com a entidade final (ou objeto) através de uma relação de determinado tipo. Na figura 1 está representado um subconjunto dos vértices (entidades) e arestas (relações) de um GC. Os métodos de aprendizagem automática em grafos aprendem uma representação vetorial das entidades no grafo denominada "embedding", de forma a que a representação no espaço reflita os seus relacionamentos com outras entidades no GC. As representações vetoriais das entidades e relações são otimizadas de modo a permitir prever os tripletos (ou factos) presentes no GC. Após o treino, as representações resultantes podem ser utilizadas para prever novos tripletos ou novos factos. Aqui demonstramos uma aplicação de métodos de aprendizagem automática em grafos a um GC biológico, relacionando entidades como genes, processos biológicos e doenças, e apresentamos a sua aplicação à previsão de associações gene-doença.

Esta figura contém uma representação gráfica de um conjunto de factos de um GC.

Os círculos (vértices) representam entidades e as ligações (arestas) representam relações entre estas. Neste caso as cores denotam o tipo da entidade. Os círculos a laranja representam genes e os círculos a azul representam doenças. Estas entidades estão ligadas por relações do tipo "IS_ASSOCIATED", representadas na figura como setas direcionadas do gene para a doença.

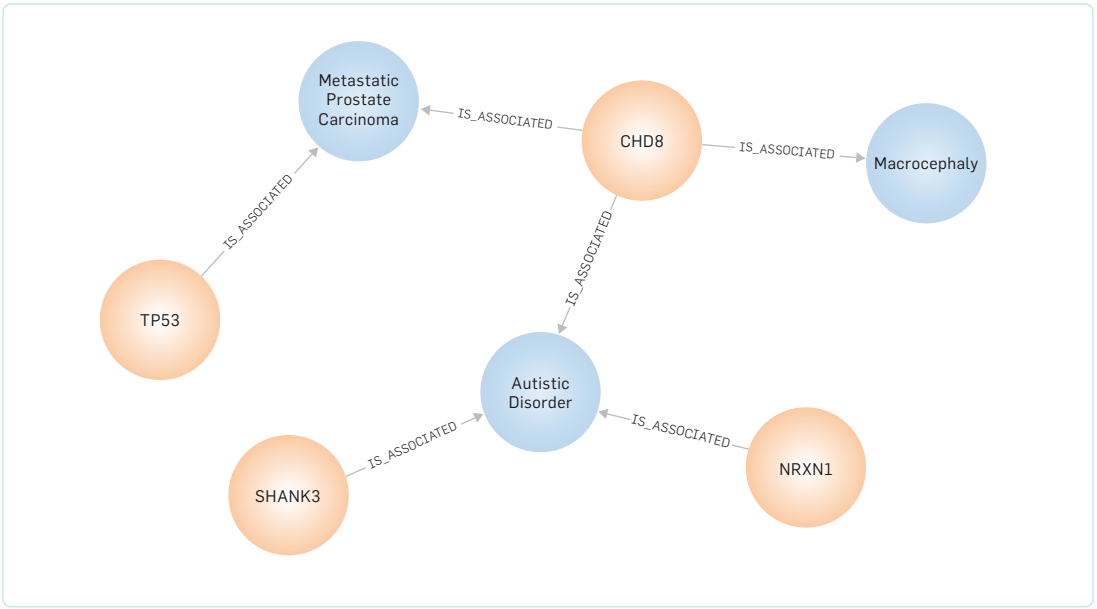
_Objetivos

O objetivo deste estudo é identificar novas associações gene-doença através do algoritmo desenvolvido e testar o seu desempenho na identificação de associações já descritas na literatura que não estejam presentes no conjunto de dados de treino do GC.

_Material e métodos

Numa primeira fase construímos um GC composto por uma série de entidades e as suas relações, extraídas de várias bases de dados biológicas. Este GC contém 7 tipos únicos de entidades: genes, doenças, fenótipos, grupos de doenças, funções moleculares, componentes celulares e processos biológicos. Os genes são representados pelos seus identificadores do Ensembl (https://ensembl.org), as doenças, fenótipos e grupos de doenças são representados por identificadores do Sistema Médico Unificado (UMLS), con-

Figura 1: Representação de um grafo de conhecimento.



forme incluído na DisGeNET (https://disgenet.org), e termos da Gene Ontology (GO) (http://geneontology.org) para processos biológicos, funções moleculares e componentes celulares, representados por seus respectivos identificadores GO. O GC tem 1.785.464 tripletos, compostos por 99.525 entidades únicas e 31 tipos de relacionamento. Essas entidades compreendem 25.450 genes, 21.623 doenças, 958 grupos de doenças, 7.409 fenótipos, 11.153 funções moleculares, 4.184 componentes celulares e 28.748 processos biológicos. O GC incorpora 901.472 associações gene-doença, 234.920 associações gene-fenótipo e 160.867 associações de gene-grupo, sendo o resto das relações anotações gene-GO (397.763) e relações entre termos da GO (90.442). Ao GC assim construído é possível aplicar algoritmos especialmente desenvolvidos para aprendizagem automática em grafos. Existem vários algoritmos para este efeito. Neste trabalho experimentamos três dos mais utilizados: ComplEx [6], DistMult [7] e TransE [8]. Utilizámos as implementações do pacote *Deep Graph Library – Knowledge Embedding* (DGL-KE) [9]. Este programa é orientado para aplicação em grafos de grande tamanho, com milhões de entidades e relações. Neste caso foi feita uma separação do GC em conjuntos de treino, teste e validação, de 80/10/10 por cento.

_Resultados

Através da previsão de novas associações gene-doença, foi possível identificar associações entre entidades do GC que não estavam presentes no conjunto de dados de treino. Calculamos os *scores* de todas estas previsões, e descartamos as associações já presentes nos dados de treino. As restantes associações são consideradas novas associações.

Destacamos na tabela 1 as 20 principais associações gene-doença previstas. Paralelamente, verificámos as evidências já existentes que ligam estes genes às doenças associadas em duas bases de dados diferentes: *Online Mendelian Inheritance in Man* (OMIM) e Orphanet. Várias das novas associações são suportadas pelos registos encontrados nestas bases de dados e são indicados na última coluna da tabela 1.

Nos casos em que não encontramos uma associação explícita do gene à doença, foi possível estabelecer uma associação entre o gene e o órgão afetado (e.x. *COL17A*1/Alopecia), entre o gene e uma via metabólica relacionada (e.x. *HTR2A*/Dependência de Opiáceos), ou uma ligação do gene a uma síndrome que inclui o fenótipo associado (*e.x. PIGN*/Microcefalia). Os genes que resultam deste tipo de evidência constituem novos candidatos a serem considerados em estudos futuros direcionados para as doenças envolvidas nestas novas associações.

Tabela 1: *Top* 20 das novas associações gene-doença com *score* mais elevado.

Ensembl ID	Gene Symbol	Disease	DB accession
ENSG00000143632	*ACTA1*	Generalized amyotrophy	OMIM:102610
ENSG00000169083	*AR*	Gastrointestinal Carcinoid Tumor	–
ENSG00000102001	*CACNA1F*	Cone Dystrophy	OMIM:300476
ENSG00000065883	*CDK13*	Byzanthine arch palate	–
ENSG00000152093	*CFC1B*	Left Atrial Isomerism	–
ENSG00000224318	*CHL1-AS2*	Scoliosis, Isolated, Suscecibility to, 3	–
ENSG00000065618	*COL17A1*	Alopecia	–
ENSG00000102468	*HTR2A*	Opiate Addiction	–
ENSG00000132740	*IGHMBP2*	Peripheral motor neuropathy	OMIM:182960
ENSG00000151704	*KCNJ1*	Metabolic alkalosis	OMIM:241200
ENSG00000163380	*LMOD3*	Pena-Shokeir syndrome type I	–
ENSG00000134571	*MYBPC3*	Tachycardia, Ventricular	ORPHA:54260
ENSG00000101247	*NDUFAF5*	Nicotinamide adenine dinucleotide coenzyme Q reductase deficiency	OMIM:618238
ENSG00000197563	*PIGN*	Strabismus	–
ENSG00000197563	*PIGN*	Microcephaly	–
ENSG00000148606	*POLR3A*	Strabismus	–
ENSG00000112619	*PRPH2*	Autosomal recessive retinitis pigmentosa	ORPHA:791
ENSG00000140522	*RLBP1*	Rod-Cone Dystrophy	OMIM:607475
ENSG00000170381	*SEMA3E*	Hirschsprung Disease	–
ENSG00000154743	*TSEN2*	Pontoneocerebellar hypoplasia	OMIM:612389

_Discussão

No *top* 20 das previsões deste estudo (tabela 1), identificámos a associação do gene *PIGN* que codifica uma proteína envolvida na biossíntese de âncoras de glicosilfosfatidilinositol (GPI) à Microcefalia e ao Estrabismo. Outras associações envolvendo doenças oculares associam as distrofias do bastonete e do cone aos genes *CACNA1F* e *RLBP1*. Outra associação importante é a do gene *AR* ao cancro gastrointestinal, gene anteriormente associado ao cancro da próstata. O gene *CDK13* aparece associado a defeitos no palato, e foi anteriormente ligado a dismorfias faciais. É também nova a associação da Alopecia ao *COL17A1*, um gene anteriormente associado a doenças epidérmicas.

O algoritmo apresentou um bom desempenho na identificação de doenças associadas a RNAs. O *CHL1-AS2* é um RNA não codificante já associado à Escoliose Idiopática e surge aqui associado à suscetibilidade à escoliose. Foram também identificados novos genes candidatos para a dependência de opiáceos (*HTR2A*) e para a síndrome de Pena-Shokeir tipo I (ORPHA: 994), uma síndrome genética rara caracterizada por diminuição dos movimentos fetais, restrição do crescimento intrauterino, contraturas articulares ou hipoplasia pulmonar. A nova associação ao gene *LMOD3* pode constituir uma pista importante para o estudo destas patologias. Outra nova associação é a da doença de Hirschsprung (OMIM: 142623), caracterizada pela perda de motilidade do trato gastrointestinal, ao gene *SEMA3E*.

_Conclusões

A metodologia aplicada neste trabalho apresentou um bom desempenho na previsão de associações gene-doença já reportadas na literatura, e identificou novas associações que estão de acordo com evidências biológicas suportadas. A previsão de novos genes associados a doenças é uma ferramenta valiosa no contexto da medicina personalizada. Estudos futuros poderão ser desenvolvidos a partir destas novas associações, envolvendo a análise de mutações em genes candidatos resultantes destas previsões, o que poderá contribuir para novos desenvolvimentos na área do diagnóstico e para a definição de intervenções terapêuticas mais adequadas.

Referências bibliográficas:

(1) Vicente AM, Ballensiefen W, Jönsson JI. How personalised medicine will transform healthcare by 2030: the ICPerMed vision. J Transl Med. 2020 Apr 28;18(1):180. https://doi.org/10.1186/s12967-020-02316-w

(2) Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. Fertil Steril. 2018 Jun;109(6):952-963. https://doi.org/10.1016/j.fertnstert.2018.05.006

(3) Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans Knowl Data Eng. 2017 Dec;29(12):2724-43. https://doi.org/10.1109/TKDE.2017.2754499

(4) Mohamed SK, Nounu A, Nováček V. Biological applications of knowledge graph embedding models. Brief Bioinform. 2021 Mar 22;22(2):1679-1693. https://doi.org/10.1093/bib/bbaa012

(5) Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020 Jun 2;18:1414-1428. https://doi.org/10.1016/j.csbj.2020.05.017

(6) Trouillon T, Welbl J, Riedel S, et al. Complex Embeddings for Simple Link Prediction. In: Balcan MF, Weinberger KQ (eds). Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research. 2016;48:2071-2080. http://proceedings.mlr.press/v48/trouillon16.html

(7) Yang B, Yih W-t., He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. (arXiv:1412.6575 [cs.CL]) (v4 29 Aug 2015). https://arxiv.org/abs/1412.6575

(8) Bordes A, Usunier N, Garcia-Duran A, et al. Translating Embeddings for Modeling Multi-relational Data. In: Burges CJC, Bottou L, Welling M, et al. (eds.) Advances in Neural Information Processing Systems. 2013;26:2787–95. http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.pdf

(9) Zheng D, Song X, Ma C, et al. DGL-KE: Training Knowledge Graph Embeddings at Scale. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York NY, Association for Computing Machinery, 2020 pp. 739–48. https://experts.umn.edu/en/publications/dgl-ke-training-knowledge-graph-embeddings-at-scale