Tecniche di Deep Learning per il Drug Design

Isabella Mendolia[1], **Salvatore Contino**[1], **Ugo Perricone**[2], **Roberto Pirrone**[1], **Edoardo Ardizzone**[1]

[1]Dipartimento di Ingegneria - Università degli Studi di Palermo,
[2]Gruppo Drug Design, Fondazione Ri.MED, Palermo

{isabella.mendolia, salvatore.contino, roberto.pirrone, edoardo.ardizzone}@unipa.it,
uperricone@fondazionerimed.com

Abstract

Il progetto di un nuovo farmaco o *Drug Design* è un processo lungo e costoso che in questi ultimi anni ha beneficiato del Machine Learning e in particolare delle Reti Neurali Profonde (DNN) per accelerare molte delle sue fasi. In questa ricerca si propone un approccio al *Virtual Screening*, che è una delle prime fasi del Drug Design, attraverso l'uso di reti convoluzionali e si riportano i primi risultati oltre alla collocazione della ricerca nel panorama nazionale e internazionale.

1 Motivazioni della ricerca

La progettazione di un nuovo farmaco è un'attività complessa che può essere considerata sia come pura ricerca scientifica sia come sviluppo industriale. Durante il processo di Drug Design si ha la necessità di consultare database pubblici o privati contenenti informazioni di natura diversa (dati chimici, biologici, tossicologici, biochimici) e non sempre facili da interconnettere. Le query che l'utente può utilizzare in tali database possono puntare dunque ad informazioni di diverso tipo e puntano principalmente a capire la compatibilità tra una molecola host e una guest coinvolte in un riconoscimento molecolare, nonché alla potenziale tossicità che questo riconoscimento può causare.

L'attività di ricerca presentata in questo lavoro utilizza reti neurali convoluzionali (CNN) per il *Virtual Screening* cioè quella fase del Drug Design in cui si individuano quelle molecole, all'interno di un database, che con maggiore probabilità risulteranno biologicamente attive su un certo target biologico, normalmente un enzima o una proteina. Recentemente la letteratura scientifica sul Drug Design ha riservato enorme attenzione al Deep Learning come paradigma per accelerare la scoperta di nuovi farmaci [Coley *et al.*, 2017; Jing *et al.*, 2018].

Il nostro caso di studio sono le chinasi. Queste proteine sono fondamentali per la regolazione, positiva o negativa, di quasi tutti i processi biochimici che avvengono durante il fisiologico avanzamento del ciclo di vita della cellula; di conseguenza riuscire ad individuare in modo efficiente quali molecole ne influenzano questa attività ha risvolti concreti nella realizzazione di farmaci migliori per il trattamento di patologie di diversa natura. [Guevara, 2016; Seo *et al.*, 2017] Gli obiettivi a lungo termine di questa ricerca sono due: creare una DNN in grado di generalizzare e predire l'attività biologica di un determinato composto rispetto ad una qualsiasi proteina e, viceversa, di predire quali composti risultano attivi rispetto ad una determinata proteina. In quanto segue si presentano le sperimentazioni condotte e l'inquadramento delle attività nel panorama della ricerca nazionale e internazionale.

2 Materiali e metodi

In questa prima fase, lo studio si è concentrato sulla sottofamiglia delle "chinasi ciclina-dipendente" (Cicline Dependent Kinase - CDK) che giocano un ruolo chiave nella regolazione del ciclo cellulare, studiando nel particolare quali molecole inibiscono l'attività della CDK1. Lo scenario indagato rappresenta un riferimento fondamentale nel drug design in ambito oncologico [Whittaker *et al.*, 2017; Mayer, 2015], tuttavia ci ha garantito in prima battuta l'uso di un dataset di dimensioni sufficientemente limitate, mirato a provare la bontà delle soluzioni proposte. I dati di attività per ciascuna molecola sono ricavati dal database *ChEMBL* [Gaulton *et al.*, 2017]. Il parametro utilizzato per distinguere tra loro molecole inibitrici e non, è lo $IC50$ ovvero la concentrazione della molecola di farmaco necessaria ad inibire il 50% dell'attività del target, nel nostro caso la CDK1.

Il livello di performance di una DNN dipende fortemente dalla buona rappresentazione dei dati di input e il problema di trovare il miglior *embedding* per descrivere la struttura molecolare 2D/3D è ancora un dibattito aperto in chemioinformatica [Hu *et al.*, 2012; Lo *et al.*, 2018]. In questo progetto si è scelto di utilizzare le *fingerprint* che sono maschere di bit utilizzate per descrivere gli atomi presenti in una molecola oltre che i loro legami e le posizioni all'interno della molecola stessa. La letteratura di settore riporta diverse proposte di fingerprint per la descrizione della molecola sia in senso bidimensionale, cioè come catena di atomi legati tra loro, sia catturando esplicitamente la struttura 3D della molecola stessa. Queste differiscono sia per quanto riguarda la lunghezza della maschera (256, 512 o 1024 bit) sia per il modo con cui le informazioni strutturali sono codificate. In generale, una fingerprint si crea attraverso l'applicazione di un kernel al vicinato di ogni atomo al fine di generare una maschera di bit, detta pattern, ossia una stringa binaria che codifica gli atomi presenti nel vicinato e i legami che li collegano. Il numero di

bit utilizzati per rappresentare un singolo pattern è in genere 4 o 5. Viene anche generata una posizione all'interno della stringa che si utilizza per posizionare il pattern tramite tecniche di hashing e combinarlo in `or` bit a bit con la maschera globale. La generazione delle fingerprint a partire dai risultati delle query su ChEMBL viene eseguita con l'utilizzo di KNIME [Berthold *et al.*, 2008], un software open-source, fra i più utilizzati in ambito chemoinformatico [Lo *et al.*, 2018].

2.1 Descrizione della rete utilizzata

L'architettura proposta nella nostra ricerca è una CNN monodimensionale che si applica alle fingerprint. In particolare, sono stati utilizzati quattro strati convoluzionali rispettivamente con 512, 256, 128 e 64 filtri con kernel di dimensione 3, intervallati ognuno da uno strato di *Max pooling*. Successivamente la rete presenta tre strati MLP con 1024, 512 e 256 neuroni rispettivamente e *Dropout* pari a 0.5 che si è dimostrato essere la migliore scelta di regolarizzazione rispetto al range di valori presente in letteratura. Lo strato di uscita è costituito da un solo neurone che usa la funzione di attivazione sigmoidale, visto il task di classificazione binaria della rete. L'addestramento è stato condotto a partire da un training set di da 1544 molecole estratte da ChEMBL di cui 772 attive e 772 inattivi rispetto alla CDK1. Visto il ridotto numero di campioni si è proceduto ad una 5-fold cross validation con 100 epoche di addestramento per volta. Un addestramento completo richiede circa 9 min. su GPU NVIDIA GeForce GTX1060 per cui non sono state usate tecniche esplicite per l'apprendimento degli iperparametri, ma si è proceduto a determinare per tentativi la capacità complessiva del modello (numero e dimensione degli strati convoluzionali e MLP, epoche di addestramento e Dropout). La rete è stata testata su sette diversi test set da 100 campioni ciascuno, che erano stati precedentemente espunti dai dati estratti da ChEMBL. In ogni test set è stata controllata espicitamente la percentuale tra molecole attive e inattive sulla CDK1. I risultati sono riportati in tabella 1.

Training		
Accuracy		Loss
0.9806		0.0355
Test		
Attivi-Inattivi	Accuracy	Loss
0% 100%	0.9570	0.2898
10% 90%	0.9500	0.3465
20% 80%	0.9600	0.2830
50% 50%	0.9100	0.5823
80% 20%	0.9000	0.6894
90% 10%	0.9000	0.5790
100% 0%	0.9695	0.0394

Tabella 1: Risultati addestramento e test

3 Collocazione della ricerca

L'attività di ricerca del gruppo è stata finanziata attraverso due borse di Dottorato di Ricerca innovativo a valere sul PON Ricerca e Innovazione 2014-2020, rispettivamente per il XXXIII e XXXIV ciclo in cui, oltre alla Fondazione Ri.MED è partner il Department of Pharmaceutical Chemistry, University of Vienna. Il gruppo ha avuto valutato positivamente un progetto PO-FESR Sicilia 2014-2020 Linea 1.1.5 e ha partecipato all'interno di un partenariato CINI alla call 14, Topic 3, del programma IMI2 sui temi del drug discovery.

Riferimenti bibliografici

[Berthold *et al.*, 2008] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, e Bernd Wiswedel. Knime: The konstanz information miner. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, e Reinhold Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 319–326, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[Coley *et al.*, 2017] Connor W. Coley, Regina Barzilay, William H. Green, Tommi S. Jaakkola, e Klavs F. Jensen. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling*, 57(8):1757–1772, August 2017.

[Gaulton *et al.*, 2017] Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, e Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, January 2017.

[Guevara, 2016] Tatiana Guevara. Evaluating the Effects of CDK Inhibitors in Ischemia–Reperfusion Injury Models. In Mar Orzáez, Mónica Sancho Medina, e Enrique Pérez-Payá, editors, *Cyclin-Dependent Kinase (CDK) Inhibitors*, volume 1336, pages 111–121. Springer New York, New York, NY, 2016.

[Hu *et al.*, 2012] Guoping Hu, Guanglin Kuang, Wen Xiao, Weihua Li, Guixia Liu, e Yun Tang. Performance Evaluation of 2d Fingerprint and 3d Shape Similarity Methods in Virtual Screening. *Journal of Chemical Information and Modeling*, 52(5):1103–1113, May 2012.

[Jing *et al.*, 2018] Yankang Jing, Yuemin Bian, Ziheng Hu, Lirong Wang, e Xiang-Qun Sean Xie. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *The AAPS journal*, 20(3):58, 2018.

[Lo *et al.*, 2018] Yu-Chen Lo, Stefano E. Rensi, Wen Torng, e Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546, August 2018.

[Mayer, 2015] Erica L. Mayer. Targeting Breast Cancer with CDK Inhibitors. *Current Oncology Reports*, 17(5), May 2015.

[Seo *et al.*, 2017] Jinsoo Seo, Oleg Kritskiy, L. Ashley Watson, Scarlett J. Barker, Dilip Dey, Waseem K. Raja, Yuan-Ta Lin, Tak Ko, Sukhee Cho, Jay Penney, M. Catarina Silva, Steven D. Sheridan, Diane Lucente, James F. Gusella, Bradford C. Dickerson, Stephen J. Haggarty, e Li-Huei Tsai. Inhibition of p25/Cdk5 Attenuates Tauopathy in Mouse and iPSC Models of Frontotemporal Dementia. *The Journal of Neuroscience*, 37(41):9917–9924, October 2017.

[Whittaker *et al.*, 2017] Steven R. Whittaker, Aurélie Mallinger, Paul Workman, e Paul A. Clarke. Inhibitors of cyclin-dependent kinases as cancer therapeutics. *Pharmacology & Therapeutics*, 173:83–105, May 2017.