ΑΡΘΡΟ ΑΝΑΣΚΟΠΗΣΗΣ

Υπολογιστική ανάπτυξη μοντέλων αξιολόγησης φαρμακοθεραπειών με χρήση αλγορίθμων μηχανικής εκμάθησης και ομικών δεδομένων

Ελένη Μπαρμπάνη, Μαρία Κορομηνά*, Γεώργιος Π. Πατρινός

Πανεπιστήμιο Πατρών, Σχολή Επιστημών Υγείας, Τμήμα Φαρμακευτικής, Εργαστήριο Φαρμακογονιδιωματικής και Εξατομικευμένης Θεραπείας, Πάτρα

ΠΕΡΙΛΗΨΗ

Στην εποχή της ψηφιακής υγείας και της τεχνητής νοημοσύνης, επιτακτική καθίσταται η ανάγκη της φαρμακευτικής βιομηχανίας για καινοτόμες και μετασχηματιστικές τεχνολογίες ανάπτυξης φαρμάκων. Οι αλγόριθμοι τεχνητής νοημοσύνης και μηχανικής εκμάθησης, αν και αργά, έχουν αναμφισβήτητα αρχίσει να επιφέρουν επανάσταση στον τομέα ανάπτυξης φαρμάκων τα τελευταία πέντε έτη. Στη συγκεκριμένη ανασκόπηση, περιγράφουμε τις πιο συχνά χρησιμοποιούμενες προσεγγίσεις μηχανικής εκμάθησης στα εργαλεία ανάπτυξης φαρμάκων και τις βάσεις ομικών δεδομένων. Αναλύουμε τις νέες υπολογιστικές προσεγγίσεις στο πεδίο της ανακάλυψης φαρμάκων στο πλαίσιο της ανάπτυξης και επαναστόχευσης φαρμάκων, αλλά και τις συνέργειες μεταξύ των ομικών επιστημών, της τεχνητής νοημοσύνης και της μηχανικής εκμάθησης. Επιπρόσθετα, παραθέτουμε μια μελλοντική προοπτική σχετικά με τους τρόπους με τους οποίους οι προσεγγίσεις της μηχανικής εκμάθησης θα είναι εφικτό να εφαρμοστούν προκειμένου όχι απλώς να επισπεύσουν την ανακάλυψη φαρμάκων αλλά και να ενισχύσουν την Ιατρική Ακριβείας με κύριο γνώμονα το όφελος των ασθενών και της δημόσιας υγείας.

ΛΕΞΕΙΣ ΕΥΡΕΤΗΡΙΟΥ: τεχνητή νοημοσύνη, μηχανική μάθηση, ανάπτυξη φαρμάκων, επανατοποθέτηση φαρμάκων, φαρμακευτική βιομηχανία

*** Αντεπιστέλλων Συγγραφέας**

Μαρία Κορομηνά, Εργαστήριο Φαρμακογονιδιωματικής και Εξατομικευμένης Θεραπείας, Τμήμα Φαρμακευτικής, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Πατρών, Πάτρα, Email: m.koromina@upnet.gr

Εισαγωγή

Η φαρμακευτική βιομηχανία, ειδικά στα πλαίσια της έρευνας και ανάπτυξης φαρμάκων, απαιτεί τη χρήση νέων τεχνολογιών προσαρμοσμένων στη σύγχρονη εποχή της ψηφιακής υγείας και της τεχνητής νοημοσύνης [1].

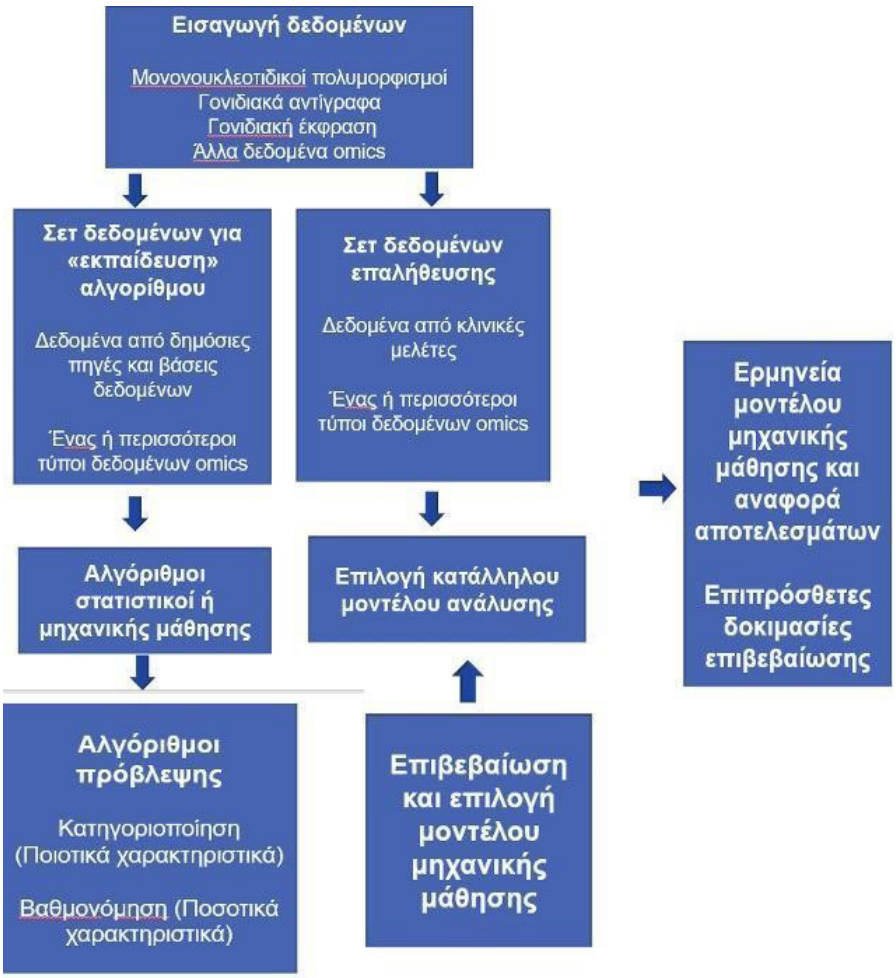
Μια σύντομη περιγραφή της ιστορίας της τεχνητής νοημοσύνης πραγματοποιήθηκε πρόσφατα από τον Garvey (2018) και διέκρινε τρία πρότυπα: το «GOFAI» (1950-60), το «Expert Systems» (τέλη 1970-80) και τη «machine learning» (2010-σήμερα). Το πρώτο πρότυπο GOFAI (Good-Old-Fashioned Artificial Intelligence), συντομογραφία της «παλιομοδίτικης τεχνητής νοημοσύνης», επικεντρώθηκε στη δημιουργία συστημάτων κοινής λογικής και οδήγησε στην ανάπτυξη θεμελιωδών τεχνικών. Το πρότυπο των «Expert systems» περιόρισε την απήχηση και κατανόησή του από το ευρύ κοινό σε ανθρώπους με εξειδίκευση σε συγκεκριμένους τομείς όπως η χημεία, η ιατρική και προσπάθησε να αναπαράγει τις γνώσεις και τις διαδικασίες λήψης των αποφάσεών τους. Το γεγονός αυτό οδήγησε στο εξειδικευμένο σύστημα τεχνητής νοημοσύνης, το MYCIN και τελικά σε πιο γνωστό λογισμικό όπως το TurboTax. Ενώ απέφεραν κάποια πρακτικά αλλά περιορισμένα αποτελέσματα, και τα δύο αυτά πρότυπα τεχνητής νοημοσύνης απέτυχαν να δημιουργήσουν τις «σκεπτόμενες μηχανές» που είχαν οραματιστεί οι εφευρέτες και πρωτοπόροι. Το τρέχον πρότυπο, η μηχανική εκμάθηση / «machine learning», έχει ξεπεράσει μερικούς από τους φραγμούς σχετικά με την ανταπόκριση και εφαρμογή της στον πραγματικό κόσμο, χάρη στην ολοένα και αυξανόμενη πληθώρα ανθρώπινων δεδομένων, στην τεράστια αύξηση της υπολογιστικής ισχύος και στην αναβίωση τόσο νευρωνικών δικτύων όσο και άλλων αλγορίθμων μηχανικής εκμάθησης. Αυτοί οι αλγόριθμοι εκμάθησης μπορούν να «εκπαιδευτούν» με σκοπό να αναγνωρίζουν και να επεκτείνουν πρότυπα προερχόμενα από ανθρώπινα δεδομένα και, επομένως, δεν απαιτούν ιδιαίτερα απαιτητικές γνώσεις προγραμματισμού [2].

Η τεχνητή νοημοσύνη και οι υπολογιστικοί αλγόριθμοι απαιτούν επανεξέταση των διαδικασιών ανάπτυξης και ανακάλυψης φαρμάκων, λαμβάνοντας παράλληλα υπόψη τις προοπτικές και τις προκλήσεις που τις συνοδεύουν. Υπάρχει η δυνατότητα ωστόσο, να αξιοποιηθούν υπό το πρίσμα της συναρπαστικής εμπειρίας των ομικών επιστημών, ώστε να συμπεριληφθεί μια ποικιλία μοριακών διαφορών μεταξύ ατόμων και πληθυσμών, η οποία μπορεί να είναι ενδεικτική της ανταπόκρισης στη θεραπεία ή των παρενεργειών, και συνεπώς να 'καθοδηγήσει' τα εργαλεία ανακάλυψης και ανάπτυξης φαρμάκων. Αυτές οι μοριακές διαφορές συχνά περιλαμβάνουν διάφορους τύπους γονιδιωματικών παραλλαγών, όπως σημειακές μεταλλάξεις, διαγραφές, παρεμβολές και μετατοπίσεις γονιδιακών αλληλουχιών, οι οποίες μπορεί να αποτελέσουν ή να υποδείξουν άμεσους μοριακούς στόχους για ανάπτυξη φαρμακευτικών θεραπευτικών προσεγγίσεων. Τέτοια παραδείγματα περιλαμβάνουν τις κλινικά ενεργές παραλλαγές, που εντοπίστηκαν στα γονίδια EGFR και ALK και μπορεί να αποτελούν στόχο των φαρμάκων της κατηγορίας αναστολέων κινάσης [3]. Οι κλινικές πληροφορίες σχετικά με τους διαφορετικούς τύπους γενετικών παραλλαγών μπορούν να εξαχθούν από κάποιες βάσεις δεδομένων όπως η ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), η COSMIC (https://cancer.sanger.ac.uk/cosmic) και η OMIM (https://www.omim.org).

Παρά την ραγδαία εξέλιξη των αλγορίθμων in silico πρόβλεψης, η υπολογιστική πρόβλεψη της φαρμακευτικής απόκρισης ειδικά σε σύνθετες και πολυπαραγοντικές ασθένειες, παραμένει μία πρόκληση. Ο μεγάλος όγκος καθώς και η ετερογένεια των δεδομένων συχνά αποτελούν τροχοπέδη στη βελτίωση της προβλεπτικής ισχύος των υπολογιστικών μοντέλων. Άλλοι κρίσιμης σημασίας προβληματισμοί κατά την ανάπτυξή τους είναι: η επιλογή των κατάλληλων συνόλων δεδομένων με στόχο την εκπαίδευση και τις δοκιμές των μοντέλων, η επιλογή των πιο ενδεδειγμένων υπολογιστικών προσεγγίσεων για εφαρμογή, καθώς και η επικύρωση και αξιολόγηση των συγκεκριμένων υπολογιστικών μοντέλων.

Η χρήση γονιδιωματικών πληροφοριών εξαγόμενων από εφαρμογές της Αλληλούχισης Επόμενης Γενιάς (Next-Generation Sequencing - NGS) χιλιάδων ασθενών σε συνδυασμό με κλινικές πληροφορίες που αφορούν τα χαρακτηριστικά γνωρίσματα της νόσου και τα θεραπευτικά αποτελέσματα, μπορεί ενδεχομένως να οδηγήσει στον εντοπισμό δεικτών σχετικών με την ανταπόκριση στη θεραπεία μέσω μιας διαδικασίας μοντελοποίησης πολλαπλών παραλλαγών. Για το σκοπό αυτό, οι εποπτευόμενοι (supervised) αλγόριθμοι μηχανικής εκμάθησης επιτρέπουν την πρόβλεψη πολλαπλών δεικτών που αφορούν τη φαρμακευτική απόκριση εφαρμόζοντας πολυ-ομικές και πολλαπλών καθηκόντων (multi-task) μεθοδολογίες εκμάθησης οι οποίες εξάγουν πληροφορίες από δείγματα ασθενών καθώς και από ομοιότητες μεταξύ των φαρμάκων [4].

Εδώ, περιγράφουμε τις πιο συχνά χρησιμοποιούμενες υπολογιστικές τεχνικές, που χρησιμοποιούν ομικά



Εικόνα 1: *Διάγραμμα ροής της ανάπτυξης υπολογιστικών μοντέλων πρόβλεψης της φαρμακευτικής απόκρισης.*

δεδομένα μεγάλης κλίμακας στις περισσότερες των περιπτώσεων. Παρέχουμε επίσης λεπτομέρειες σχετικά με τις μεθόδους της μηχανικής εκμάθησης που χρησιμοποιούνται κατά τη διαδικασία ανάπτυξης και διαλογής φαρμάκων. Τέλος, παρατίθεται και μία περιγραφή των διαθέσιμων προσεγγίσεων μηχανικής εκμάθησης που αφορούν την επαναστόχευση φαρμάκων (Drug repositioning).

Επισκόπηση της στρατηγικής σχεδιασμού υπολογιστικών μοντέλων πρόβλεψης

Στις περισσότερες περιπτώσεις, η ανάπτυξη υπολογιστικών μοντέλων πρόβλεψης της φαρμακευτικής απόκρισης βασίζεται σε τέσσερα διαφορετικά στάδια. Κατά το πρώτο στάδιο, αρχικά επιλέγονται τα κατάλληλα σύνολα δεδομένων και προ- επεξεργάζονται μέσω επιλογής των σχετικών υποσυνόλων δεδομένων. Στη συνέχεια πραγματοποιείται "κανονικοποίησή" των δεδομένων αυτών και φιλτράρισμα, εξάλειψη των δεδομένων που αντιστοιχούν σε θόρυβο καθώς και εκείνων που περιέχουν πληροφορίες μη σχετικές με την εκάστοτε μελέτη (**Εικόνα 1**) [5]. Αυτά τα σύνολα δεδομένων μπορεί να αποτελούνται από μονονουκλεοτιδικούς πολυμορφισμούς, παραλλαγές του πλήθους των αντιγράφων των γονιδίων (Copy Number Variants, CNVs) και από δεδομένα που προκύπτουν από τη γονιδιακή έκφραση. Ενδιαφέρον παρουσιάζει το εύρημα συγκριτικών μελετών ανάλυσης ότι τα δεδομένα γονιδιακής έκφρασης έχουν και την πιο ισχυρή προγνωστική αξία, ενώ τα υπολογιστικά μοντέλα αύξησαν οριακά την ακρίβεια πρόβλεψης της φαρμακευτικής απόκρισης [6].

Η πιθανότητα εμφάνισης μη αναμενόμενων αποκρίσεων κατά το πλαίσιο των θεραπειών είναι απαραίτητο να ληφθεί υπόψη κατά την εφαρμογή αλγορίθμων με στόχο την εκτίμηση της ευαισθησίας σε ορισμένη φαρμακευτική αγωγή. Για το λόγο αυτό, τα σύγχρονα μοντέ-

Πίνακας 1.					
Δημόσια Βάση δεδομένων	***NCI-DREAM7***	***CCLE***	***NCI-60***	***GDSC***	***TCGA***
Καρκινικοί τύποι	*Καρκίνος μαστού*	36 τύποι καρκίνου	9 τύποι καρκίνου	29 τύποι καρκίνου	33 τύποι καρκίνου
Αριθμός δειγμάτων	*53 κυτταρικές σειρές*	947 κυτταρικές σειρές	*59 κυτταρικές σειρές*	*1124 κυτταρικές σειρές*	11000 δείγματα ασθενών

Πίνακας 1: *Συνοπτική καταγραφή των δημοσίων βάσεων ποικίλων ομικών δεδομένων μαζί με τον αντίστοιχο καρκινικό τύπο και τον αριθμό δειγμάτων κάθε βάσης (CCLE: Cancer Cell Line Encyclopedia, DREAM7: Dialogue on Reverse Engineering Assessment and Methods, GDSC: Genomics of Drug Sensitivity in Cancer, NCI: National Cancer Institute , TCGA: The Cancer Genome Atlas).*

λα ενσωματώνουν δεδομένα από ένα ευρύτερο φάσμα κυτταρικών σειρών ή/και κλινικών δειγμάτων ώστε να επιτευχθεί ακριβέστερη εκτίμηση της κλινικής ετερογένειας ή/και των αποκρίσεων στη θεραπεία [4].

Το δεύτερο στάδιο σχεδιασμού του αλγορίθμου περιλαμβάνει τη φάση εκπαίδευσης του μοντέλου που έχει επιλεχθεί (**Εικόνα 1**). Μπορεί έτσι να εφαρμοστεί μια ποικιλία τεχνικών μηχανικής εκμάθησης ενώ με τα δεδομένα μπορούν να τροφοδοτηθούν διαφορετικά, εναλλακτικά μοντέλα ώστε να επιλεγεί το ακριβέστερο σε πρόβλεψη και να εξελιχθεί. Ο τύπος των δεδομένων καταχώρησης και τα ειδικά χαρακτηριστικά γνωρίσματα του ζητήματος πρόβλεψης της φαρμακευτικής απόκρισης απαιτούν προσεκτική εξέταση. Το τρίτο βήμα σχεδιασμού του υπολογιστικού μοντέλου πρόβλεψης αναφέρεται συχνά ως ανεξάρτητη αξιολόγηση. Κατά τη διάρκεια αυτού, διεξάγονται πολλαπλές δοκιμές σε ανεξάρτητα μεταξύ τους σύνολα δεδομένων από το επιλεγμένο και εξελιγμένο μοντέλο. Αυτό αποσκοπεί να επαληθεύσει ότι το υποψήφιο μοντέλο πρόβλεψης μπορεί με ακρίβεια να προβλέψει τις φαρμακευτικές αποκρίσεις πάνω σε νεοφανή, μη προηγουμένως μελετημένα δεδομένα που προέρχονται από διαφορετικά εργαστήρια και πλατφόρμες μέτρησης.

Τελικό στάδιο αποτελεί η εφαρμογή του μοντέλου σε δεδομένα που σε κλινικό επίπεδο μοιάζουν με τα χαρακτηριστικά της υπό εξέταση ασθένειας. Για παράδειγμα, μοντέλα που εκπαιδεύονται με τη χρήση δεδομένων προερχόμενων από κυτταρικές σειρές μπορούν να δοκιμαστούν σε στερεές ή υγρές βιοψίες ασθενών. Και πάλι, η ερμηνεία και η εφαρμογή των κατάλληλων μεθοδολογιών για την αποσαφήνιση των προβλέψεων καθίστανται κρίσιμες παράμετροι (**Εικόνα 1**).

Αξιολόγηση του μοντέλου πρόβλεψης

Η αξιολόγηση του μοντέλου αποτελεί κρίσιμο και αναπόσπαστο κομμάτι κατά τη διαδικασία δημιουργίας ενός ισχυρού και ακριβούς μοντέλου in silico πρόβλεψης, καθώς αυτό θα πρέπει να έχει ικανοποιητική απόδοση ακόμη, και ιδίως, σε δεδομένα που δεν έχει 'δει' ποτέ. Η αξιολόγηση του μοντέλου πραγματοποιείται αφού 'εκπαιδευτεί' ο αλγόριθμος με ένα σύνολο δεδομένων εκπαίδευσης (training set), ενώ έπεται η επαλήθευση του μοντέλου με τη χρήση ανεξάρτητων συνόλων δεδομένων (test set), ώστε να εκτιμηθεί το τελικό μοντέλο. Μεταξύ των πιο σημαντικών παραμέτρων απόδοσης που προορίζονται για εφαρμογή σε ζητήματα κατηγοριοποίησης δεδομένων είναι: (α) η συνολική ευαισθησία (Sensitivity) του μοντέλου, (β) η ακρίβεια (Precision), (γ) η ανάκληση (Recall), (δ) η περιοχή κάτω από την καμπύλη (Area Under Curve). Επιπλέον αξιολογούνται και οι καμπύλες ακριβείας-ανάκλησης (Precision-Recall curves) που αποτελούν παράγωγο μέγεθος των (β) και (γ).

Η συνήθης διαδικασία αξιολόγησης του προσαρμοζόμενου μοντέλου είναι γνωστή ως Διασταυρούμενη Επικύρωση (Cross-Validation, CV). Κατά τη διάρκεια αυτής, το αρχικό σύνολο δεδομένων χωρίζεται σε δύο ξεχωριστά υποσύνολα, εκ των οποίων το ένα θα χρησιμοποιηθεί για την εκπαίδευση του αλγορίθμου και το άλλο για τον έλεγχο της απόδοσης του μοντέλου. Τα πιο συχνά χρησιμοποιούμενα σχήματα διασταυρούμενης επικύρωσης είναι η 'K-fold' (K-Fold Cross Validation, KF-CV) και η 'Leave-one-out' (Leave-One-Out Cross Validation, LOO-CV) διασταυρούμενη επικύρωση. Κατά τη μέθοδο KF-CV, το αρχικό σύνολο δεδομένων κατανέμεται σε μέρη K, με τα μέρη K-1 να χρησιμοποιούνται για εκπαί-

δευση και το τελευταίο για δοκιμές επαλήθευσης του εκάστοτε μοντέλου. Η αρχή λειτουργίας της μεθόδου LOO-CV προϋποθέτει ότι ένα μεμονωμένο υποσύνολο από το αρχικό σύνολο δεδομένων θα περιλαμβάνει το σύνολο των εναπομείναντων δεδομένων και το οποίο θα προορίζεται για επαλήθευση. Στη συνέχεια, η διαδικασία επαναλαμβάνεται τόσες φορές όσα και τα εξεταζόμενα δεδομένα (data points) [7], [8].

Στην εικόνα 2 αναπαρίστανται συνοπτικά τα στάδια ανάπτυξης υπολογιστικών μοντέλων πρόβλεψης της φαρμακευτικής απόκρισης:

Δημόσιες πηγές ομικών δεδομένων για τη δημιουργία αλγορίθμων πρόβλεψης της απόκρισης σε φαρμακευτική θεραπεία

Όπως αναλύθηκε στην προηγούμενη ενότητα, η πλειοψηφία των μοντέλων πρόβλεψης της φαρμακευτικής απόκρισης εκπαιδεύεται με τη χρήση συνόλων δεδομένων που δημιουργούνται από διαφορετικά ερευνητικά προγράμματα. Παρόλο που αυτή η προσέγγιση μπορεί να έχει βιολογική ακρίβεια, υπόκειται σε ένα σύνολο περιορισμών, όπως είναι ο αριθμός των ασθενών που υπόκεινται σε ανάλυση ή οι ψευδείς αναγνώσεις ομικών και κλινικών δεδομένων τα οποία μπορούν ενδεχομένως να ενσωματωθούν στην ανάπτυξη μοντέλων. Μία εναλλακτική προσέγγιση, η οποία κερδίζει ολοένα και μεγαλύτερη αποδοχή, είναι η εκπαίδευση (ή/και δοκιμή) μοντέλων σε δεδομένα δημοσίων βάσεων δεδομένων προερχόμενα από μεγάλα ερευνητικά προγράμματα (**Πίνακας 1**).

Κλασσικές προσεγγίσεις μηχανικής εκμάθησης στην πρόβλεψη της φαρμακευτικής απόκρισης

Γενικά, οι τεχνικές μηχανικής εκμάθησης κατηγοριοποιούνται στις εποπτευόμενες (supervised), οι οποίες λαμβάνουν υπόψη πληροφορίες σχετικά με τις κατηγορίες των δεδομένων εκπαίδευσης (training data) και στις μη εποπτευόμενες (unsupervised), οι οποίες στοχεύουν στη δημιουργία ομάδων μεταξύ των δεδομένων εκπαίδευσης. Όσον αφορά τις τελευταίες, η ένταξη των δεδομένων (data points) σε μία ομάδα γίνεται με τέτοιο τρόπο ώστε τα στοιχεία μιας ομάδας να είναι όσο το δυνατόν πιο όμοια μεταξύ τους και να διαφέρουν όσο το δυνατό περισσότερο από εκείνα άλλων ομάδων. Συνήθως, ανάλυση που βασίζεται στις μη εποπτευόμενες μεθόδους προηγείται και είναι δυνατό να χρησιμοποιηθεί ώστε να αποκτήσουμε μία πρώτη εκτίμηση των δεδομένων [9], [10], [11].

Οι εποπτευόμενες μέθοδοι μηχανικής εκμάθησης περιλαμβάνουν ένα ευρύ φάσμα τεχνικών, το οποίο χωρίζεται σε δύο κύριες κατηγορίες. Η κύρια κατηγορία περιλαμβάνει τους αλγόριθμους κατηγοριοποίησης (classification algorithms) και παλινδρόμησης (regression-based algorithms) με τους πρώτους να χρησιμοποιούνται για τον προσδιορισμό της κατηγορίας ένταξης μιας νέας καταχώρησης, για παράδειγμα αν μια καρκινική κυτταρική σειρά αναμένεται να ανταποκριθεί με τον επιθυμητό ή με ανεπιθύμητο τρόπο μετά από τη χορήγηση ενός συγκεκριμένου φαρμάκου [12], [13]. Η δεύτερη κατηγορία τεχνικών αποσκοπεί στην εκτίμηση της αξίας μιας εξεταζόμενης μεταβλητής [14], [15]. Από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους κατηγοριοποίησης είναι οι επονομαζόμενοι ως Support Vector Machines (SVMs) και τα Random Forests [16], [17], [18], [19], [20]. Ωστόσο, βελτιωμένη προγνωστική απόδοση είναι δυνατό να επιτευχθεί υπό διάφορες συνθήκες [13] χρησιμοποιώντας μοντέλα που βασίζονται στη μέθοδο της παλινδρόμησης, όπως τα elastic net ή ridge regression [21].

Προσεγγίσεις μηχανικής εκμάθησης στην επαναστόχευση φαρμάκων

Η επαναστόχευση φαρμάκων (Drug repositioning) ορίζεται ως η διαδικασία επιλογής ενός ήδη γνωστού φαρμάκου για μια εναλλακτική φαρμακολογική χρήση. Λόγω της ραγδαίας ανάπτυξης του τομέα της βιοπληροφορικής και της ανάλυσης ομικών δεδομένων μεγάλης κλίμακας, η επαναστόχευση φαρμάκων έχει μειώσει σημαντικά τον απαιτούμενο χρόνο κατά τη διαδικασία ανάπτυξης φαρμάκων. Στις μέρες μας, οι ερευνητές χρειάζονται περίπου 1 με 2 χρόνια μόνο για να εντοπίσουν νέους πιθανούς στόχους φαρμάκων και περίπου 8 έτη ώστε να αναπτύξουν κατά προσέγγιση ένα επαναστοχευμένο φάρμακο (**Εικόνα 2**).

Συνοπτικά, οι προσεγγίσεις επαναστόχευσης φαρμάκων μπορούν να ταξινομηθούν σε τρεις κατηγορίες: προσεγγίσεις βάσει δικτύου (network based approaches), προσεγγίσεις εξόρυξης κειμένου (text-mining approaches) και σημασιολογικές προσεγγίσεις (semantic approaches). Οι μέθοδοι της πρώτης κατηγορίας (network-based methods) διαιρούνται περαιτέρω στις μεθόδους ομαδοποίησης (clustering) και διάδοσης (propagation). Οι πρώτες αναζητούν τυχόν συσχετίσεις μεταξύ φαρμάκων και νόσων ή πιθανών στόχων σε μικρότερες επιμέρους "συστάδες" ενός μεγαλύτερου δικτύου, ενώ οι τελευταίες βασίζονται σε πρωτύτερα

Εικόνα 2: *Διάγραμμα ροής της σταδιακής διαδικασίας που ακολουθείται κατά την επαναστόχευση φαρμάκων.*

ανακτημένες συσχετίσεις μέσω των αντίστοιχα χρησιμοποιουμένων δικτύων.

Τέλος, περαιτέρω διάκριση των ανωτέρω μεθόδων μπορεί να γίνει με κριτήριο τη σύνθεση των δημιουργούμενων δικτύων. Πιο συγκεκριμένα, αυτά που σχηματίζονται χρησιμοποιώντας έναν τύπο πληροφοριών, όπως οι πρωτεϊνικές αλληλεπιδράσεις (Protein-Protein Interactions-PPIs) χαρακτηρίζονται ως ομοιογενείς, ενώ δίκτυα που ενσωματώνουν διάφορους τύπους δεδομένων, όπως αυτά που προέρχονται από αναλύσεις ομικών δεδομένων είναι γνωστά ως ετερογενή [22].

Αντίθετα, η μέθοδος εξόρυξης κειμένου (text-mining approach) εκμεταλλεύεται το τεράστιο εύρος της διαθέσιμης βιβλιογραφίας, το οποίο στη συνέχεια φιλτράρεται για τη διατήρηση μόνο εκείνων των πηγών δεδομένων που θα επιτρέψουν την εξαγωγή πληροφοριών για τους υπό μελέτη βιολογικούς όρους. Τέλος, στις σημασιολογικές προσεγγίσεις (semantic approaches), η δημιουργία δικτύων καθοδηγείται από προηγούμενες βιοϊατρικές γνώσεις, όπως εξάγονται από βάσεις δεδομένων και στη συνέχεια χρησιμοποιούνται αλγόριθμοι μηχανικής εκμάθησης για τον εντοπισμό νέων αλληλεπιδράσεων και σχέσεων, οι οποίες υπάρχουν σε αυτά τα δίκτυα [23].

Συζήτηση και μελλοντικές προοπτικές

Στις μέρες μας, όλα τα στάδια της ανακάλυψης και ανάπτυξης φαρμάκων, από την επιλογή και επικύρωση του στόχου μέχρι και τις κλινικές δοκιμές μπορούν να υπόκεινται και να επωφελούνται από την εφαρμογή αλγορίθμων και εργαλείων μηχανικής εκμάθησης. Μεταξύ των διαφορετικών προκλήσεων που μπορούν να αντιμετωπιστούν μέσω αυτών των μεθοδολογιών περιλαμβάνονται τόσο η απαιτητική διαδικασία αναγνώρισης νέων μοριακών στόχων όσο και η απόκτηση μιας βαθύτερης γνώσης τόσο των μηχανισμών της νόσου, αλλά και των πολύπλοκων και πολυπαραγοντικών φαινοτύπων πολλών ασθενειών [24],[25].

Οι προσεγγίσεις της τεχνητής νοημοσύνης (που συμπεριλαμβάνει και τη μηχανική εκμάθηση) μπορούν να ωφελήσουν ουσιαστικά κάθε βήμα της διαδικασίας σχεδιασμού και ανάπτυξης φαρμάκων. Ξεκινώντας από τη διαδικασία διαλογής (screening) χημικών ενώσεων μέσα από τεράστιες βιβλιοθήκες και συνεχίζοντας με την ανακάλυψη της πρόδρομης ένωσης και την ταυτοποίηση της ένωσης οδηγού [26], [27], έως και τη διερεύνηση των πιθανών τροποποιήσεων που θα οδηγήσουν σε ενώσεις με βελτιωμένα χαρακτηριστικά φαρμακοκινητικής, φαρμακοδυναμικής και τοξικολογίας. Επιπροσθέτως, βελτιστοποιείται και το στάδιο σχεδιασμού της συνθετικής τους πορείας συνήθως μέσω αναδρομικής σύνθεσης [28], [29]. Οι αλγόριθμοι μηχανικής εκμάθησης μπορούν να χρησιμοποιηθούν επίσης και για τον εντοπισμό νέων βιοδεικτών με στόχο τη βελτίωση της αποτελεσματικότητας των φαρμάκων, ενισχύοντας έτσι τον τομέα της Ιατρικής Ακριβείας [30].

Επιπλέον, μια πληθώρα μοντέλων μηχανικής εκμάθησης για την ανακάλυψη βιοδεικτών και για την πρόβλεψη της φαρμακευτικής ευαισθησίας υπόσχεται να βελτιώσει σημαντικά τα ποσοστά της κλινικής επιτυχίας, τόσο μέσω της συμβολής τους στην αποκάλυψη μοριακών μηχανισμών δράσεων, αλλά και μέσω της

παροχής των απαιτούμενων γνώσεων που προϋποθέτει η επίτευξη της εξατομίκευση της θεραπείας [31], [32]. Αυτό επιτυγχάνεται πρωτίστως μέσω της ανάλυσης, της διερεύνησης και της ερμηνείας των ομικών δεδομένων με τη χρήση μεθοδολογιών τεχνητής νοημοσύνης και μηχανικής εκμάθησης. Περαιτέρω εφαρμογή των μεθόδων μηχανικής εκμάθησης πραγματοποιείται στους τομείς της φαρμακευτικής βιομηχανίας όπως στον τομέα της χημειο-πληροφορικής, της γονιδιωματικής πληροφορικής και της βιοϊατρικής απεικόνισης [33], [34], [35].

Η εισαγωγή αλγορίθμων και εργαλείων μηχανικής εκμάθησης καθίσταται εφικτή λόγω της αυξημένης υπολογιστικής ισχύος σε συνδυασμό με την αυξημένη διαθεσιμότητα σε δεδομένα μεγάλου όγκου, με διάφορες φαρμακευτικές εταιρείες να επενδύουν τελευταία σε αυτή. Ωστόσο, παρά τις υψηλές προσδοκίες των αλγορίθμων/εργαλείων αυτών στο πεδίο της ανακάλυψης φαρμάκων και της ιατρικής ακριβείας, υπάρχουν μόνο ορισμένες περιπτώσεις όπου βιοδείκτες και μοντέλα πρόβλεψης έχουν εφαρμοστεί σε κλινικές δοκιμές. Σημαντικοί παράγοντες, που επηρεάζουν την υιοθέτηση τέτοιων προγνωστικών μοντέλων, είναι η επιλογή των κατάλληλων μοντέλων, η δυνατότητα αναπαραγωγής μοντέλων που δημιουργούνται με τη χρήση νευρωνικών δικτύων, η πρόσβαση σε επιμελημένα δεδομένα καθώς και ο σχεδιασμός κατάλληλων δοκιμών για κλινικό περιβάλλον [29] .

Μεταξύ των πιο σημαντικών περιορισμών που πρέπει να εξεταστούν προσεκτικά στον τομέα εφαρμογής της μηχανικής εκμάθησης στην ανακάλυψη και στην επαναστόχευση φαρμάκων, είναι η ποιότητα των πειραματικών δεδομένων. Ένα μοντέλο μπορεί να είναι τόσο ακριβές όσο και τα δεδομένα εκπαίδευσης αυτού. Εάν τα δεδομένα είναι εσφαλμένα, εάν δημιουργούν 'θόρυβο' και εάν η συλλογή και η ενσωμάτωσή τους δεν ακολουθεί μια συστηματική και καλά χαρακτηρισμένη διαδικασία αξιολόγησης αυτών, τότε είναι πολύ πιθανό ότι η πρόβλεψη του αλγορίθμου θα είναι ανακριβής. Εάν ο αλγόριθμος και οι χρήστες του δεν λάβουν υπόψη αυτές τις παρατηρήσεις, τότε τα αποτελέσματα της πρόβλεψης μέσω μηχανικής εκμάθησης θα είναι επίσης εσφαλμένα [29].

Χωρίς αμφιβολία, η διαδικασία μηχανικής εκμάθησης παραμένει ένα «μαύρο κουτί» [36]. Από την στιγμή που οι λειτουργίες του μοντέλου δεν προσδιορίζονται ρητά, ο δημιουργός του αλγορίθμου μπορεί να μην γνωρίζει τι ελέγχεται κατά τη διάρκεια των ενδιάμεσων σταδίων ή και την ακριβή διαδικασία που οδηγεί στο συγκεκριμένο αποτέλεσμα/πρόβλεψη του αλγορίθμου. Για το σκοπό αυτό, η πρόσληψη του κατάλληλα εκπαιδευμένου προσωπικού που θα αποσκοπεί στην γεφύρωση του χάσματος μεταξύ των ομικών επιστημών και της τεχνητής νοημοσύνης, είναι ένας άλλος παράγοντας που πρέπει να εξεταστεί προσεκτικά.

Ανεξάρτητα από τους παραπάνω περιορισμούς, η επίτευξη συνεργασίας μεταξύ φαρμακευτικών εταιρειών και εταιρειών που επικεντρώνονται στην ανάπτυξη προσεγγίσεων μηχανικής εκμάθησης, θα μπορούσε να βοηθήσει στην ταυτοποίηση νέων θεραπευτικών μορίων, αλλά και στην ανακάλυψη νέων θεραπευτικών μεθόδων μέσω της χρήσης δεδομένων μεγάλου όγκου, με απώτερο σκοπό τη βελτιστοποίηση της υγείας. Λαμβάνοντας υπόψη το υψηλό χρονικό αντίτιμο, τους πόρους και την προσπάθεια σχεδιασμού και ανάπτυξης νέων φαρμάκων, η οποία συνοδεύεται από μειωμένα ποσοστά έγκρισης, εύλογα στηριζόμαστε σε υπολογιστικές προσεγγίσεις που αποσκοπούν στην περαιτέρω βελτίωση της αποτελεσματικότητας της διαδικασίας ανάπτυξης φαρμάκων. Αυτές οι εξελίξεις θα συμβάλλουν στη βελτίωση των υπηρεσιών υγειονομικής περίθαλψης, αλλά και στην επίτευξη της ιατρικής ακριβείας [28].

Συμπεράσματα

Συνοπτικά, περιγράψαμε τη λογική και τα βήματα σχεδιασμού αλγορίθμων μηχανικής εκμάθησης, οι οποίοι μπορούν να χρησιμοποιηθούν κατά την μέθοδο της in silico διαλογής φαρμάκων (in silico drug screening). Οι αλγόριθμοι αυτοί επιτρέπουν τον εντοπισμό δραστικών ουσιών, των οποίων η αναστολή ή η ενίσχυση είναι δυνατό να βελτιώσει την αποτελεσματικότητα των θεραπευτικών προσεγγίσεων. Επιπλέον, όταν εφαρμοστεί η μέθοδος εξόρυξης δεδομένων (data-mining), τότε είναι δυνατό να αποσαφηνιστούν σημαντικές σχέσεις μεταξύ γονιδίων και φαρμακευτικής απόκρισης. Για το σκοπό αυτό, παρέχουμε επίσης μία σύνοψη χρήσιμων προσεγγίσεων μηχανικής εκμάθησης που χρησιμοποιούνται στην επαναστόχευση φαρμάκων, η οποία αποτελεί μια εναλλακτική προσέγγιση της 'παραδοσιακής' μεθόδου ανακάλυψης φαρμάκων. Εν κατακλείδι, η τρέχουσα ανασκόπηση αποδεικνύει ότι το πεδίο της Τεχνητής Νοημοσύνης, που περιλαμβάνει τη μηχανική εκμάθηση, μπορεί να εφαρμοστεί στην κλινική πράξη με κύριο σκοπό τη βελτίωση της Ιατρικής Ακριβείας και της διαδικασίας ανακάλυψης φαρμάκων/διαλογής φαρμάκων.

ABSTRACT

Computational development of drug response prediction models via machine learning approaches and omics data

Eleni Barbani, Maria Koromina, George P. Patrinos

University of Patras, School of Health Sciences, Department of Pharmacy, Laboratory of Pharmacogenomics and Individualized Therapy, Patras, Greece

In the age of digital health and artificial intelligence, the pharmaceutical industry needs innovative and transformative drug development technologies. Undoubtedly, artificial intelligence and machine learning algorithms have begun to revolutionize the drug development industry over the past five years, although at a slow pace. In this review, we describe the most commonly used "machine learning" approaches in drug development tools and databases. We also analyse new computational approaches in the field of drug discovery and namely in the context of drug repurposing, as well as the synergies between group sciences, artificial intelligence and machine learning. In addition, we present a future perspective on the ways in which machine learning approaches can be applied, in order not only to accelerate the discovery of drugs, but also to enhance Precision Medicine with the most benefit for patients and public health.

KEY WORDS: artificial intelligence, machine learning, drug development, drug repurposing, pharmaceutical industry

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Kohane, I. S. Ten things we have to do to achieve precision medicine. Science, 2015;349, 37.
2. Garvey, C. Interview with Colin Garvey, Rensselaer Polytechnic Institute. Artificial Intelligence and Systems Medicine Convergence. OMICS. 2018; 22, 130-132.
3. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., JR. & Kinzler, K. W. Cancer genome landscapes. Science. 2013; 339, 1546-1558.
4. Azuaje, F. Computational models for predicting drug responses in cancer research. Brief Bioinform. 2017; 18, 820-829.
5. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16, 321-332.
6. Costello, J. C., Heiser, L. M., Georgii, E., et al. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechn. 2014;32, 1202-1212.
7. Baek, S., Tsai, C. A. & Chen, J. J. Development of biomarker classifiers from high-dimensional data. Brief Bioinform. 2019;10, 537-546.
8. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. BMC bioinformatics. 2016;7, 91-91.
9. Byers, L. A., Diao, L., Wang, J., et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin Cancer Res. 2017; 19, 279-290.
10. Nicolau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc Natl Acad Sci U S A. 2011;108, 7265-7270.
11. Gholami, A.M., Hahne, H., Wu, Z., et al. Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Rep. 2013; 4, 609-620.
12. Fersini, E., Messina, E. & Archetti, F. A p-Median approach for predicting drug response in tumour cells. BMC Bioinformatics. 2014; 15.
13. Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H. & Margolin, A. A. Systematic assessment of analytical methods

for drug sensitivity prediction from cancer cell line data. Pacific Symposium on Biocomputing. Pac Symp Biocomput. 2014;63-74.
14. Falgreen, S., Dybkaer, K., Young, K. H., et al. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. BMC Cancer. 2015;15.
15. Neto, E. C., Jang, I. S., Friend, S. H. & Margolin, A. A. The Stream algorithm: computationally efficient ridge-regression via Bayesian model averaging, and applications to pharmacogenomic prediction of cancer cell line sensitivity. Pac Symp Biocomput. 2014; 27-38.
16. Amin, S. B., Yip, W. K., Minvielle, S., et al. Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. Leukemia. 2014; 28, 2229-2234.
17. Chen, B., Sirota, M., Fan-Minogue, H., Hadley, D. & Butte, A. J. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Med Genomics. 2015;8 Suppl 2, S5-S5
18. Cortés-Ciriano, I., Van Westen, G. J. P., Bouvier, G., et al. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. Bioinformatics. 2015;32, 85-95.
19. Tran, T. P., Ong, E., Hodges, A. P., Paternostro, G. & Piermarocchi, C. Prediction of kinase inhibitor response using activity profiling, in vitro screening, and elastic net regression. BMC Syst Biol. 2014;8, 74-74.
20. Stetson, L. C., Pearl, T., Chen, Y. & Barnholtz-Sloan, J. S. Computational identification of multi-omic correlates of anticancer therapeutic response. BMC Genomics. 2014;15, S2.
21. Consortium, C. C. L. E. C. G. O. D. S. I. C. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015;528, 84-87.
22. Wu, Z., Wang, Y. & Chen, L. (2013). Network-based drug repositioning. Mol Biosyst. 2013;9, 1268.
23. Xue, H., Li, J., Xie, H. & Wang, Y. Review of Drug Repositioning Approaches and Resources. Int J Biol Sci. 2018;14, 1232-1244.
24. Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W. & Zhang, X. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics. 2017;33, 2010-2019.
25. Ferrero, E., Dunham, I. & Sanseau, P. In silico prediction of novel therapeutic targets using gene-disease association data. Journal of translational medicine. J Transl Med. 2017; 15, 182-182.
26. Anderson, A. C. Structure-based functional design of drugs: from target to lead compound. Methods Mol Biol. 2012; 823, 359-366.
27. Zhu, T., Cao, S., Su, P.-C., et al. Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based on a Critical Literature Analysis. J Med Chem. 2013; 56, 6560-6572.
28. Mak, K.-K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today. 2019;24, 773-780.
29. Vamathevan, J., Clark, D., Czodrowski, P., et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019; 18, 463-477.
30. Mamoshina, P., Volosnikova, M., Ozerov, I. V., et al. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Front Genet. 2018; 9, 242-242.
31. Li, B., Shin, H., Gulbekyan, G., et al. Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib. PloS one. 2015; 10, e0130700-e0130700.
32. Kraus, V. B. Biomarkers as drug development tools: discovery, validation, qualification and use. Nat Rev Rheumatol. 2018; 14, 354-362.
33. Gonczarek, A., Tomczak, J. M., Zaręba, S., Kaczmar, J., Dąbrowski, P. & Walczak, M. J. Interaction prediction in structure-based virtual screening using deep learning. Comput Biol Med. 2018; 100, 253-258.
34. Ekins, S. The Next Era: Deep Learning in Pharmaceutical Research. Pharm Res. 2016; 33, 2594-2603.
35. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. Mol Syst Biol. 2016; 12, 878-878.
36. Lamberti, M. J., Wilkinson, M., Donzanti, B. A., et al. A Study on the Application and Use of Artificial Intelligence to Support Drug Development. Clin Ther. 2019; 41, 1414-1426.