Δεδομενοθηρία ή θεωριολαγνεία;

Στάθης Ψύλλος
ΕΚΠΑ

Αλλαγή μεθοδολογικού παραδείγματος στην μοριακή βιολογία. Η αλλαγή αυτή προκαλείται από μια έκρηξη στην συλλογή αξιόπιστων δεδομένων. Το νέο αναδυόμενο μεθοδολογικό παράδειγμα, in sillico discovery, υπόσχεται την ανίχνευση προτύπων (patterns) στα δεδομένα, χρησιμοποιώντας, μεταξύ άλλων, υπολογιστικές τεχνικές πάνω σε μεγάλες ποσότητες δεδομένων. Πρόκειται, θα έλεγε κανείς, για μια μετάβαση από τον έλεγχο θεωριών μέσω δεδομένων στην *εξαγωγή* της θεωρίας από τα δεδομένα. Υπό μία έννοια, τίθεται στο προσκήνιο η διαμάχη μεταξύ υποθετικο-παραγωγής και επαγωγής. Αλλά αυτό δεν είναι ακριβές.

I

Ποια είναι η σχέση δεδομένων και θεωρίας; Τα δεδομένα παίζουν ένα *διπλό* ρόλο. Πρώτον, θέτουν περιορισμούς στον λογικό χώρο των θεωριών. Οι προτεινόμενες θεωρίες πρέπει, κατ' ελάχιστον, να είναι συμβατές με τα δεδομένα. Δεύτερον, ελέγχουν τις προτεινόμενες θεωρίες—είτε επικυρώνοντάς τις είτε διαψεύδοντας τις. Άρα, τα δεδομένα λειτουργούν ως *φίλτρα* στο χώρο των θεωριών.

Σημειώστε ότι η πληθώρα των δεδομένων όντως παίζει κάποιον ιδιαίτερο ρόλο στην παραπάνω διαδικασία. Αν κάποιος είναι επαγωγιστής, τότε όσο πιο πολλά είναι τα δεδομένα που συνηγορούν υπέρ μια υπόθεσης, τόσο μεγαλύτερος ο βαθμός επικύρωσης της. Αλλά η πληθώρα των δεδομένων δεν εκφράζεται, κατ' ανάγκην, μέσω ενός πλούσιου περιεχομένου της θεωρίας. (πχ. Όλοι οι κόρακες είναι μαύροι.) Αν κάποιος είναι διαψευσιοκράτης, αυτό που μετρά δεν είναι η ποσότητα την δεδομένων αλλά η ποιότητά τους. Ένας δυνάμει διαψευστής αρκεί. Φυσικά όσο περισσότερους δυνάμει διαψευστές έχει μια θεωρία, τόσο πιο πλούσιο είναι το περιεχόμενό της. Αλλά, από λογική άποψη, ένας αρκεί. Συνεπώς υπάρχει μια ασυμμετρία. Για τον επαγωγιστή, η πληθώρα των δεδομένων είναι επιστημικά σημαντική, αλλά για τον διαψευσιοκράτη δεν είναι.

Σε κάθε περίπτωση, υπάρχουν ανεξάρτητοι λόγοι για να επερωτήσουμε το αίτημα της πληθώρας των δεδομένων. Αυτό που εν τέλει μετρά δεν είναι η ποσότητα των

δεδομένων, αλλά η ποιότητά τους. Πχ. το πόσο ποικίλα είναι, το πόσο αυστηρά ελέγχουν τη θεωρία. (πχ. Νεύτωνας) Η συλλογή περαιτέρω ποιοτικά ομοίων δεδομένων δεν θα οδηγήσει κανένα στο να πάρει ένα βραβείο Νόμπελ. (Σημειώστε ότι το προλεχθέν δεν οδηγεί στη θέση ότι η διαψευσιοκρατία είναι καλύτερη από τον επαγωγισμό. Το θέμα είναι ότι ακόμα και ο επαγωγιστής έχει να κερδίσει από την ποιότητα και όχι την ποσότητα των δεδομένων. Αν η θεωρία επιβεβαιώνεται σε πεδία στα οποία ήταν πιθανόν τα δεδομένα να την αντικρούουν, τόσο το καλύτερο για τη θεωρία).

Έτσι η αλλιώς, τα δύο μοντέλα επιστημονικής μεθόδου αντιμετωπίζουν σημαντικά εννοιολογικά προβλήματα.

Επαγωγή: τι σημαίνει συλλέγω δεδομένα; Η θεωρητική φόρτιση της παρατήρησης. Επίσης, ποιες αρχές διέπουν την εν λόγω συναγωγή; (αντιπροσωπευτικότητα δείγματος κλπ.)

Υποθετικο-παραγωγή: οι θεωρίες ελέγχονται ως σύνολα και ως εκ τούτου δεν είναι διαψεύσιμες, αυστηρά μιλώντας. (πχ Ποσειδώνας, Ερμής) Το πρόβλημα των Duhem-Quine και ο ποιοτικός (μη αλγοριθμικός) χαρακτήρας της επιστημονικής μεθόδου. Επίσης, το πρόβλημα των εναλλακτικών υποθέσεων-εξηγήσεων.

Ο σαφής προσδιορισμός της επιστημονικής μεθόδου είναι ακόμα ένα ανοικτό θέμα.

II

Η ανωτέρω ίσως είναι μια θεωριο-κεντρική προσέγγιση στη μέθοδο. Αλλά πια είναι η πιθανή αντίπαλη αντίληψη;

Οι εξελίξεις στη σύγχρονη μοριακή βιολογία, κυρίως μέσω του προγράμματος χαρτογράφησης του γονιδιώματος, τείνουν να προτείνουν μια, ας πούμε, δεδομενο-κεντρική προσέγγιση. Γιατί να μην αφήσουμε τα δεδομένα να μιλήσουν από μόνα τους, ιδιαιτέρως αφού έχουμε, ή χωρίς μεγάλο κόστος μπορούμε να έχουμε, μια πληθώρα αυτών;

Τι σημαίνει όμως ‘να αφήσουμε τα δεδομένα να μιλήσουν από μόνα τους’; Αντί τα δεδομένα να λειτουργήσουν ως φίλτρα υπαρχουσών θεωριών ή υποθέσεων να λειτουργήσουν ως οι *εξαγωγείς* θεωριών ή υποθέσεων. Η ιδέα είναι ότι εάν τα δεδομένα είναι αρκετά και αξιόπιστα, τότε οι θεωρίες μπορούν να *εξαχθούν* από αυτά. Οι θεωρίες γίνονται περιπτώσεις ή συμπυκνώσεις των δεδομένων.

Η ιδέα είναι παλιά. Francis Bacon. Αλλά εξίσου παλιό είναι και το πρόβλημα που αντιμετωπίζει: δεν χρειαζόμαστε τις θεωρίες απλώς για να περιγράψουμε τα δεδομένα αλλά και για να *εξηγήσουμε*. Ο εξηγητικός ρόλος των θεωριών δεν μπορεί ούτε να αναχθεί στα, ούτε να εξαχθεί από, τα δεδομένα. Η θεωρία υπερβαίνει τα δεδομένα (και άρα δεν μπορεί να εξαχθεί από αυτά) ακριβώς γιατί τα εξηγεί.

Η καινούργια έκφραση της παλιάς ιδέας στηρίζεται στην έννοια του *προτύπου*. Ο στόχος είναι να βρούμε πρότυπα τα οποία είναι παρόντα στα δεδομένα. Η έννοια του προτύπου είναι πολύ χρήσιμη. Κατ' αρχήν, σημειώστε ότι η ανωτέρω περιγραφή του επιστημονικού παιγνίου δεν έρχεται σε αντίθεση με όσα προανέφερα. Μπορεί να πει κανείς ότι το επιστημονικό παίγνιο συνίσταται δε δυο βήματα (όχι κατ' ανάγκην υπό την ακόλουθη χρονική σειρά): ανακάλυψη προτύπων στα δεδομένα και εξήγησή τους μέσω θεωρητικών υποθέσεων. Αν η έμφαση στα πρότυπα τείνει απλώς να τονίσει ότι η θεωρία δεν πρέπει να αγνοεί τα πρότυπα στα δεδομένα ή ότι τα πρότυπα στα δεδομένα δεν πρέπει να διαμορφώνονται ή παραποιούνται έτσι ώστε να δικαιώνεται μια προκαταβολικά διατυπωμένη θεωρία, τότε η έμφαση αυτή είναι μεθοδολογικά αθώα, απολύτως αληθής και καλοδεχούμενη. Φοβούμαι όμως ότι η έμφαση στα πρότυπα εκλαμβάνεται είτε ως μια οιονεί-αλγοριθμική διαδικασίας ανακάλυψης θεωριών, είτε ως *αντικατάσταση* του ρόλου των θεωριών (μέσω της εξίσωσης της θεωρίας με το πρότυπο). Μια τέτοια αντίληψη είναι λανθασμένη.

Να μερικοί (από τους πάρα πολλούς) λόγους γι' αυτό. Πρώτον, η έννοια του προτύπου είναι, εν μέρη, αξιολογική. Το τι συνιστά ένα πρότυπο εξαρτάται, εν μέρη, από διάφορες παραδοχές σε σχέση με την ομοιομορφία, αρμονία κπλ. Δεύτερον, ας υποθέσουμε ότι η έννοια του προτύπου δεν είναι, ούτε εν μέρη, αξιολογική αλλά ότι στηρίζεται σε μια έννοια αντικειμενική έννοια ομοιότητας. Όπως είναι γνωστό, οποιοδήποτε πράγμα είναι όμοιο με οποιοδήποτε άλλο, εκτός και εάν θέσουμε κάποιους περιορισμούς στο βαθμό και στις πλευρές της ομοιότητας. Αυτό όμως σημαίνει ότι η ανίχνευση προτύπων στα δεδομένα προϋποθέτει μια θεωρητική κατανόηση (έστω και υποτυπώδη) του τι συμβαίνει σε αυτά, του ποιες ιδιότητες είναι σχετικές κλπ. Τρίτον, όπως έχει φανεί από το λεγόμενο Inductive Learning στην επιστήμη των υπολογιστών, η διαδικασία ανακάλυψης προτύπων στα δεδομένα προϋποθέτει την πρότερη σύλληψη κάποιων τύπων *συναρτησιακών μορφών* τις οποίες θα πάρουν αυτά τα πρότυπα. Χωρίς μια τέτοια σύλληψη (που αν μη τι άλλο περιχαρακώνει τη *μορφή* της θεωρίας που θα εξαχθεί από τα δεδομένα) τα δεδομένα δεν μπορούν να τιθασευτούν. Τέταρτον, η ανακάλυψη ενός προτύπου στα δεδομένα

προϋποθέτει ότι τα δεδομένα μπορούν να αξιολογηθούν θεωρητικά ως προς την σημασία τους. Υπό μια έννοια, όσο πιο πολλά και πλούσια είναι τα δεδομένα, τόσο πιο πολλά πρότυπα μπορούν να ανιχνευθούν σε αυτά. Το ποια εξ αυτών των προτύπων είναι σημαντικά ή έχουν θεωρητικές συνέπειες είναι ένα κατ' εξοχήν θεωρητικό πρόβλημα ερμηνείας.

Η θεωρία εμπλέκεται με δύο τρόπους στο πρόβλημα που συζητούμαι. Ο ένας είναι με τη μορφή υποθέσεων και παραδοχών που συνιστούν τμήμα της ίδιας της εμπειρικής μεθοδολογίας και του πειράματος. Υπό αυτή την έννοια, η θεωρία είναι *μέσο* για την ανίχνευση προτύπων στα δεδομένα. Ο άλλος είναι με τη μορφή *τελικού προϊόντος*, δηλ. ως το αποτέλεσμα της μεθοδολογικής επεξεργασίας των δεδομένων. Νομίζω ότι οι οπαδοί της in sillico discovery έχουν δίκιο να τονίζουν ότι το τελικό προϊόν περιορίζεται πιο δραστικά εάν τα δεδομένα είναι πλείστα και πλούσια (αν και θα έλεγα ότι αυτό που μετρά είναι η ποιοτική και όχι η ποσοτική ποικιλομορφία τους). Όμως, μάλλον έχουν άδικο εάν τονίζουν ότι το τελικό προϊόν προκύπτει από τα δεδομένα με έναν τρόπο ελεύθερο θεωρίας. Δεν είναι παράδοξο η ίδια η θεωρία να είναι μέσο *και* προϊόν. Αυτό είναι σύνηθες στη φυσική, όταν για παράδειγμα κάποιες υψηλού επιπέδου υποθέσεις επικουρούμενες από δεδομένα οδηγούν στην εξαγωγή μεσαίου επιπέδου υποθέσεων. Αυτή είναι η μέθοδος που ο Νεύτων ονόμασε «παραγωγή από τα φαινόμενα».

Θεωρήστε ένα πείραμα ο οποίο ανιχνεύει αλλαγές στην έκφραση ενός γονιδίου κατά την διάρκεια κάποιας αναπτυξιακής αλλαγής ενός οργανισμού. Ας υποθέσουμε ότι ανιχνεύουμε 1053 γονίδιο στα οποία το επίπεδο του mRNA ανεβαίνει, 393 στα οποία κατεβαίνει και 4126 στα οποία παραμένει ως είχε. Είναι φανερό ότι το πώς αυτά τα δεδομένα θα αποτιμηθούν, και το τι είδους πρότυπα θα διαφανούν, εξαρτάται από μια σειρά μεθοδολογικές και θεωρητικές υποθέσεις υποβάθρου. Αυτές είναι αναπόδραστες. Μόνο μέσω αυτών τα δεδομένα «αποκτούν φωνή».

Θα ήταν λάθος, πιστεύω, να πει κάποιος ότι η θεωρία συναντά τα δεδομένα μόνο στη φάση που τίθεται υπό τον έλεγχό τους, δηλ. μόνο στη φάση της επιβεβαίωσης ή της διάψευσής της. Αντιθέτως, η θεωρία εμποτίζει την επιστημονική δραστηριότητα (ρητά ή άρρητα) σε όλες τις εκφάνσεις της. Πολλές φορές η συλλογή δεδομένων δεν αποσκοπεί στον έλεγχο μιας θεωρίας. Άλλες φορές η θεωρία δεν είναι ικανοποιητικά αναπτυγμένη για να ελεγχθεί και η συλλογή και αποτίμηση των δεδομένων βοηθά στην αποκρυστάλλωση της. Αλλά από αυτά δεν έπεται ότι η θεωρία δεν εμπλέκεται

στην συλλογή και αποτίμηση δεδομένων. Για να παραφράσω τον Κλάουζεβιτς, η αλίευση δεδομένων είναι η συνέχιση της θεωρίας με άλλα μέσα.