Κεφάλαιο 1

Εισαγωγή στη Βιοπληροφορική

Σύνοψη
Η Βιοπληροφορική είναι ένας ταχέα αναπτυσσόμενος διεπιστημονικός κλάδος. Παρόλο που ένας ακριβής ορισμός δεν μπορεί να δοθεί, και υπάρχουν μάλιστα και πολλές διαφωνίες ανάλογα με την οπτική και το υπόβαθρο του καθενός, είναι σαφές ότι πρόκειται για τον επιστημονικό κλάδο που βρίσκεται στην περιοχή επαφής της βιολογίας με τα μαθηματικά και την επιστήμη υπολογιστών. Στο κεφάλαιο αυτό, θα προσπαθήσουμε να εξετάσουμε τέτοια θέματα από όλες τις πλευρές. Θα δούμε το ιστορικό πλαίσιο ανάπτυξης της βιοπληροφορικής (ή καλύτερα, της υπολογιστικής βιολογίας), το διεπιστημονικό χαρακτήρα της, τους μύθους που τη συνοδεύουν, αλλά θα δούμε και τις τελευταίες εξελίξεις στη βιβλιογραφία της βιοπληροφορικής, τόσο διεθνώς όσο και στην Ελλάδα. Με τα περιεχόμενα αυτού το κεφαλαίου, ευελπιστούμε ότι οι αναγνώστες θα μπορέσουν να αποκτήσουν μια εποπτική εικόνα αυτού του σύνθετου ερευνητικού πεδίου η οποία θα τους βοηθήσει στην κατανόηση των επόμενων κεφαλαίων.

Προαπαιτούμενη γνώση
Ο αναγνώστης πρέπει να έχει τις βασικές γνώσεις μοριακής βιολογίας και γενετικής.

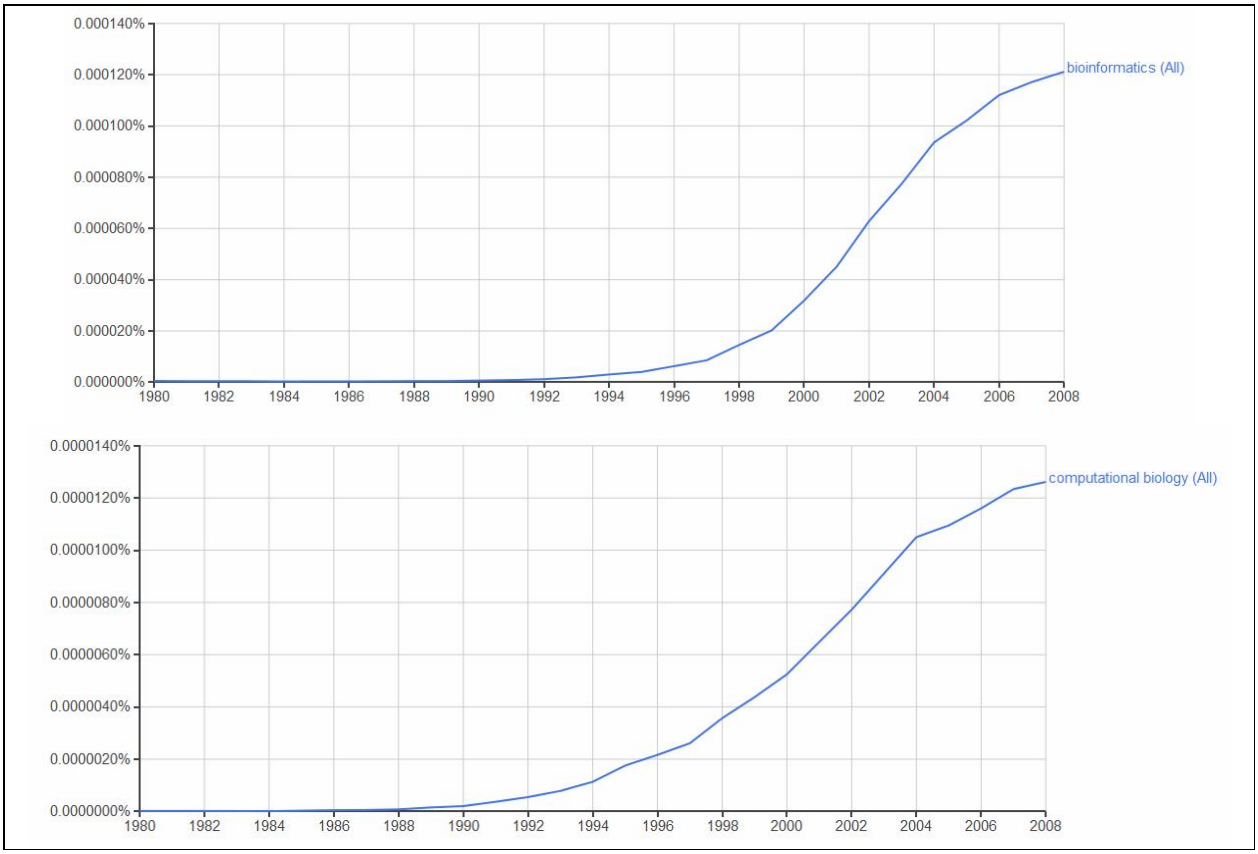
1. Εισαγωγή

Η Βιοπληροφορική είναι ένας διεπιστημονικός κλάδος, και παρόλο που ένας κοινώς αποδεκτός ορισμός δεν υπάρχει, μια προσπάθεια ορισμού της θα ήταν ως ο επιστημονικός χώρος όπου η σύμπραξη της Βιολογίας με την Πληροφορική, τη Στατιστική και τα Μαθηματικά εξερευνά νέους τρόπους για την προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της Βιολογίας. Πρόκειται για γνωστικό χώρο με συγκεκριμένο όσο και ευρύ πεδίο εφαρμογών και αλληλεπίδρασης με τη σύγχρονη δομική, μοριακή, πληθυσμιακή και περιβαλλοντική βιολογία. Ο κλάδος της Βιοπληροφορικής σήμερα θεωρείται, παγκόσμια, ένας από τους πλέον αναπτυσσόμενους, ενώ έχει ήδη επιδείξει σημαντικά επιτεύγματα και έχει συγκεντρώσει ιδιαίτερα σημαντικές επενδύσεις.

Καθώς η Βιοπληροφορική είναι διεπιστημονικός κλάδος, υπάρχουν πολλές και αντικρουόμενες απόψεις σχετικά με τον ορισμό της αλλά και σχετικά με το επιστημολογικό της καθεστώς. Η πιο απλοϊκή προσέγγιση λέει ότι η Βιοπληροφορική είναι απλώς η εφαρμογή κάποιων τεχνολογιών (μαθηματικών, υπολογιστικών, κ.ο.κ.) σε βιολογικά προβλήματα. Η πιο σύνθετη, στην οποία προσχωρεί και ο συγγραφέας αυτού του βιβλίου, είναι ότι η Βιοπληροφορική είναι πλέον μια ξεχωριστή επιστήμη, η οποία ναι μεν χρησιμοποιεί υπολογιστικά και μαθηματικά εργαλεία σε βιολογικά προβλήματα, αλλά κάνει και κάτι άλλο: παράγει (ή τουλάχιστον, προσπαθεί να παράγει) και γενικότερους νόμους που διέπουν αυτά τα βιολογικά συστήματα. Με αυτόν τον τρόπο, μιλάμε για μια βιολογικής κατεύθυνσης επιστήμη ή ειδικότητα, με τη δική της παράδοση και τις δικές της μεθοδολογίες. Πολλές φορές χρησιμοποιείται παράλληλα και ο όρος Υπολογιστική Βιολογία, ενώ από πολλούς οι δυο αυτοί όροι χρησιμοποιούνται αδιάκριτα μεταξύ τους. Όπως θα δούμε παρακάτω η άποψη αυτή μάλλον δικαιώνεται ιστορικά. Παρόλα αυτά, ένας λογικός διαχωρισμός είναι ότι ο όρος βιοπληροφορική αναφέρεται κυρίως στην πρακτική εφαρμογή, δηλαδή στη χρήση αλγόριθμων και υπολογιστικών τεχνικών που επιτρέπουν την απάντηση βιολογικών ερωτημάτων (π.χ. αναζήτηση μιας ακολουθίας σε μια βάση δεδομένων, χειρισμός μεγάλου όγκου ακολουθιών ή δεδομένων γονιδιακής έκφρασης κλπ), ενώ ο όρος υπολογιστική βιολογία είναι κάπως πιο θεωρητικός και αναφέρεται στα θεωρητικά αποτελέσματα στα οποία στηριζόμαστε για να αναπτύξουμε έναν αλγόριθμο, μια μεθοδολογία ή ένα γενικό νόμο.

Η προσωπική άποψη του συγγραφέα, είναι ότι παρόλες τις επιμέρους διαφορές που αναλύθηκαν παραπάνω, ο όρος Υπολογιστική Βιολογία θα έπρεπε να χρησιμοποιείται γενικά αντί της Βιοπληροφορικής. Και τούτο, γιατί με αυτόν τον τρόπο θα δίναμε έμφαση στο αντικείμενο που μελετάμε (τα βιολογικά συστήματα) και όχι στον τρόπο (την υπολογιστική μεθοδολογία). Με λίγα λόγια, πρέπει να γίνει κατανοητό ότι η Βιοπληροφορική/Υπολογιστική Βιολογία, είναι πρώτα από όλα Βιολογία, μελέτη των ζωντανών οργανισμών με υπολογιστικές μεθοδολογίες. Αυτό δεν σημαίνει όμως ότι πρέπει να υπαχθεί στη συντεχνιακή αντίληψη (κάτι συνηθισμένο στη χώρα μας) ότι αυτή είναι μια δραστηριότητα μόνο για Βιολόγους. Το

αντίθετο, είναι ένας διεπιστημονικός κλάδος, στον οποίο μπορούν και πρέπει να συνεισφέρουν επιστήμονες εκπαιδευμένοι σε διάφορες ειδικότητες (βιολόγοι, μαθηματικοί, επιστήμονες Η/Υ, μηχανικοί κ.ο.κ.). Αυτό όμως που πρέπει να γίνει, είναι ότι πρέπει επιπλέον, να υπάρξει και διεπιστημονική εκπαίδευση έτσι ώστε να υπάρξει ένας κοινός τόπος και μια κοινή γλώσσα στην οποία όλοι αυτοί οι ειδικοί θα μπορούν να συνεννοούνται. Και φυσικά, πρέπει να υπάρξει και προσπάθεια δημιουργίας διεπιστημονικών ατόμων, όχι μόνο ομάδων. Αξίζει να αναφερθεί πάντως, ότι παρόλο που υπάρχουν δεκάδες επιστημονικά περιοδικά με σαφή αναφορά στη Βιοπληροφορική, οι γενικότερες ταξινομήσεις των επιστημονικών περιοδικών από την ISI, περιλαμβάνουν σαν ξεχωριστή κατηγορία μόνο την γενικότερη περίπτωση (Mathematical and Computational Biology), κατηγορία που περιλαμβάνει και περιοδικά Βιοπληροφορικής και Υπολογιστικής Βιολογίας αλλά και περιοδικά Βιοστατιστικής και Ιατρικής Πληροφορικής. Όπως θα δούμε παρακάτω, τέτοιες συνέργειες με άλλα παρεμφερή επιστημονικά πεδία, είναι αρκετά κοινές στο χώρο.



Εικόνα 1.1: *Εικόνα από το google trends για τους όρους «bioinformatics» και «computational biology», αντίστοιχα.*

Στις επόμενες παραγράφους, θα προσπαθήσουμε να αποδώσουμε μια σύντομη ιστορική αναδρομή του επιστημονικού κλάδου της βιοπληροφορικής και να ανιχνεύσουμε τις ρίζες του. Θα δούμε τη διεπιστημονικότητα του, αλλά και τις περιοχές επαφής με τις γειτονικές επιστήμες και, τέλος, θα δούμε τις τάσεις στη διεθνή βιβλιογραφία αλλά και αναλυτικά την κατάσταση στην Ελλάδα.

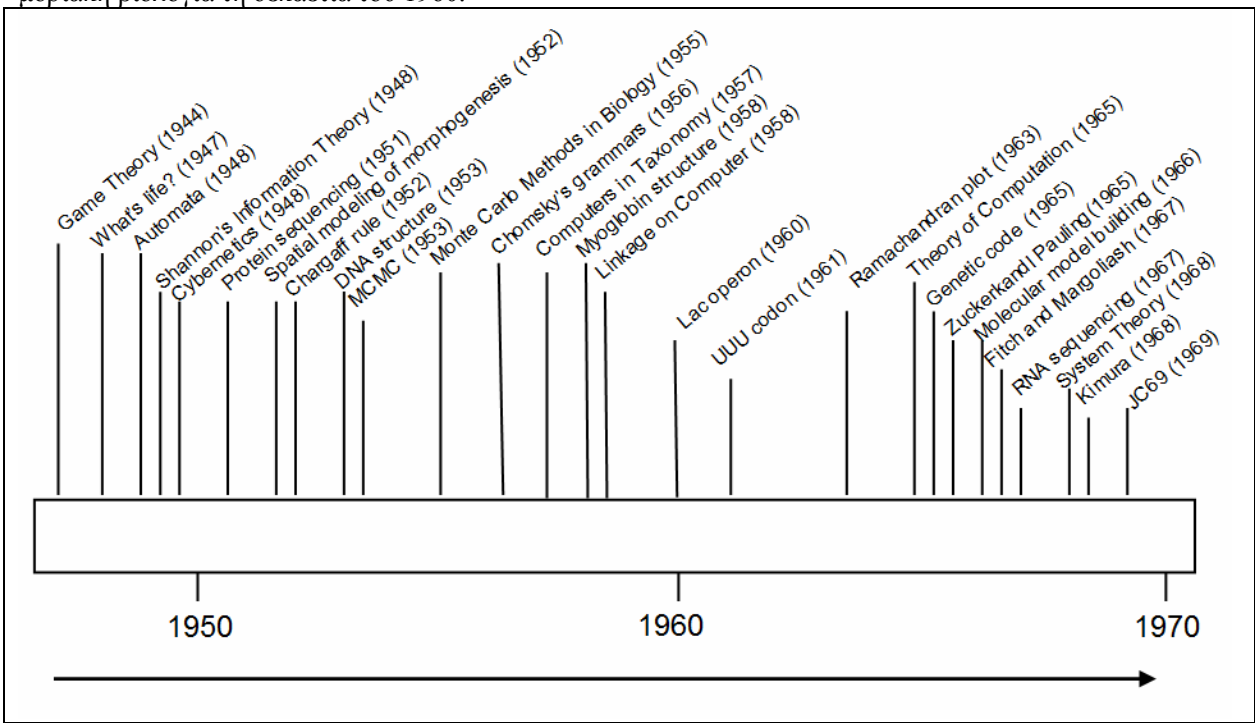
1.1. Η ιστορία της βιοπληροφορικής και της υπολογιστικής βιολογίας

Όπως είδαμε παραπάνω, ο όρος Βιοπληροφορική (Bioinformatics), είναι ιδιαίτερα πρόσφατος και εμφανίστηκε στη δεκαετία του 1990. Επίσης, είναι ένας ιδιαίτερα επιτυχημένος όρος καθώς έχει επικρατήσει η χρήση του διεθνώς, κάτι μάλλον περίεργο, ιδιαίτερα αν σκεφτούμε ότι ο όρος informatics στα Αγγλικά δεν χρησιμοποιείται και πολύ. Όπως θα δούμε παρακάτω όμως, ο όρος αυτός είναι και παραπλανητικός, καθώς συνήθως αναφερόμαστε στην υπολογιστική ανάλυση των βιολογικών συστημάτων (με βασική αναφορά στις βιολογικές αλληλουχίες) και μια τέτοια προσπάθεια δεν ξεκίνησε προφανώς τη δεκαετία του 1990, αλλά πολύ πιο πριν. Αν δούμε τα πράγματα από μια ιστορική σκοπιά, θα δούμε ότι ήδη από τις αρχές του 20^{ου} αιώνα,

υπήρχαν πολλές προσπάθειες μαθηματικοποίησης και ποσοτικοποίησης των βιολογικών φαινομένων, αλλά αυτές οι προσπάθειες συμβάδιζαν πάντα με τις υπολογιστικές μεθοδολογίες τις εποχής, όσο και με το είδος των βιολογικών δεδομένων που ήταν κάθε φορά διαθέσιμα. Τα περισσότερα από όσα περιγράφονται παρακάτω, βασίζονται σε γνωστές ιστορικές ανασκοπήσεις (Hagen, 2000; C. A. Ouzounis & Valencia, 2003; Roberts, 2000; Searls, 2010; Trifonov, 2000), αλλά και στην προσωπική εμπειρία του συγγραφέα. Προφανώς μια τέτοια θεώρηση, και ειδικά στην έκταση και την ανάλυση ενός διδακτικού εγχειριδίου, θα είναι αναγκαστικά ελλειπής, αλλά ελπίζω ότι η γενική εικόνα που θα μείνει τελικά στον αναγνώστη θα είναι αποκαλυπτική όσο και χρήσιμη.

1.1.1. Οι δεκαετίες του 1950 και 1960

Αν προσπεράσουμε τις προσπάθειες μαθηματικοποίησης της γενετικής, που έδωσαν γένεση στη γενετική πληθυσμών και τη σύγχρονη εξελικτική θεωρία, από την εποχή των Hardy, Weinberg και Fisher, Wright κλπ, θα πρέπει να ανιχνευσουμε τις απαρχές της σύγχρονης υπολογιστικής βιολογίας, στις απαρχές της ίδιας της σύγχρονης μοριακής βιολογίας τη δεκαετία του 1950 και 1960 (Εικόνα 1.2). Για παράδειγμα, τα πειράματα του Chargaff που έδειξαν ότι το ποσοστό Αδενίνης είναι το ίδιο με το ποσοστό της Θυμίνης και το ποσοστό Γουανίνης ίσο με αυτό της Κυτοσίνης σε κάθε μόριο DNA, ήταν οι πρώτες ενδείξεις για κάποια μορφή ψηφιακής πληροφορίας στις βιολογικές αλληλουχίες. Τα πειράματα αυτά, ως γνωστόν χρησιμοποιήθηκαν από τους Watson και Crick για να μπορέσουν να προσδιορίσουν την τρισδιάστατη δομή του DNA η οποία τους έδωσε και το νόμπελ (χρησιμοποιόντας δεδομενα του Wilkins, ο οποίος βραβεύτηκε μαζί τους αλλά και της Franklin η οποία όμως είχε πεθάνει στο ενδιάμεσο). Τη δεκαετία του 1960 έγιναν επίσης και οι πρωτοποριακές μελέτες των Jacob και Monod στη γονιδιακή ρύθμιση (lac operon). Ενώ όσον αφορά τις πρωτεΐνες, μετά τον προσδιορισμό των πρώτων τρισδιάστατων δομών (ινσουλίνη και μυογλοβίνη), και τη βράβευση των Perutz και Kendrew με το νόμπελ το 1962, ακολούθησαν τη δεκαετία του 1960 μια σειρά παρόμοιες σημαντικές δομές (λυσοζύμη, παπαΐνη, ριβονουκλεάση κ.ο.κ.), και άνοιξε ο δρόμος για τη μελέτη της δομής και της λειτουργίας των πρωτεϊνών σε ατομικό επίπεδο. Επίσης, η εύρεση της πρωτοταγούς δομής των πρωτεϊνών έγινε το 1951, και του RNA το 1967. Βλέπουμε λοιπόν ότι πολλά από τα προβλήματα που απασχολούν τη Βιοπληροφορική μέχρι σήμερα, έχουν τις ρίζες τους στην έκρηξη που πραγματοποιήθηκε στη μοριακή βιολογία τη δεκαετία του 1960.



Εικόνα 1.2: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληφορική μέχρι και το τέλος της δεκαετίας του 1960*

Παράλληλα, από τη δεκαετία του 1950 και του 1960 είχαν τεθεί ήδη και τα θεμέλια της σύγχρονης θεωρητικής πληροφορικής, με τη θεωρία υπολογισμού, τη θεωρία πληροφορίας του Shannon, τη μηχανή του Turing, τα αυτόματα και τη θεωρία παιγνίων του von Neumann, τη μελέτη των συμβολοσειρών (strings), την θεωρία συστημάτων, την κυβερνετική και τον ορισμό των γραμματικών από τον Chomsky. Έτσι, δεν είναι περίεργο, αν αναλογιστούμε και τα παραπάνω, ότι οι πρώτες προσπάθειες υπολογιστικής αντιμετώπισης βιολογικών προβλημάτων, εμφανίστηκαν τη δεκαετία του 1960 και σε αυτές βρίσκονται τα πρώτα ψήγματα αυτού που σήμερα ονομάζουμε Υπολογιστική Βιολογία και Βιοπληροφορική. Έτσι, η αποκρυπτογράφηση του γενετικού κώδικα, ήταν κομβικό σημείο στην ανάπτυξη της μοριακής βιολογίας και όλων των βιοεπιστημών. Αυτή η ίδια η φύση του γενετικού κώδικα, ο οποίος στην πραγματικότητα είναι μια συνάρτηση, μια διμελής απεικόνιση από το σύνολο των 64 τριπλετών στο σύνολο των 20 αμινοξέων, ήταν ήδη αντικείμενο έντονης θεωρητικής αλλά και υπολογιστικής επεξεργασίας από τη δεκαετία του 1960, ενώ όταν διαλευκάνθηκε πειραματικά έγιναν και πολλές θεωρητικές μελέτες για τις ιδιότητές του και την προέλευση του. Εμφανίστηκαν επίσης οι εφαρμογές των πρώτων υπολογιστών της εποχής στη βιολογία, με τη χρήση τους μεταξύ άλλων στην Ταξινομική και στην κατασκευή μοριακών μοντέλων για την κρυσταλλογραφία. Την ίδια εποχή, είδαμε και την πρώτη προσπάθεια χρήσης βιολογικών αλληλουχιών για εξελικτικές μελέτες από τους Zuckerkandl και Pauling, τη χρήση τους για την κατασκευή φυλογενετικών δέντρων από τους Fitch and Margoliash αλλά και τα πρώτα μαθηματικά μοντέλα της μοριακής εξέλιξης από τους Kimura και Nei. Στο επίπεδο των πρωτεϊνών είδαμε τις πρώτες εργασίες του Ramachandran για τη μελέτη των δομικών ιδιοτήτων και των περιορισμών των αμινοξικών καταλοίπων σε μια πρωτεϊνική δομή, από τις οποίες έχει προκύψει το πασίγνωστο διάγραμμα Ramachandran και οι επιτρεπτές δίεδρες γωνίες που εμφανίζονται σε πρωτεϊνικές δομές, αλλά και τα πρώτα helical wheel plots.

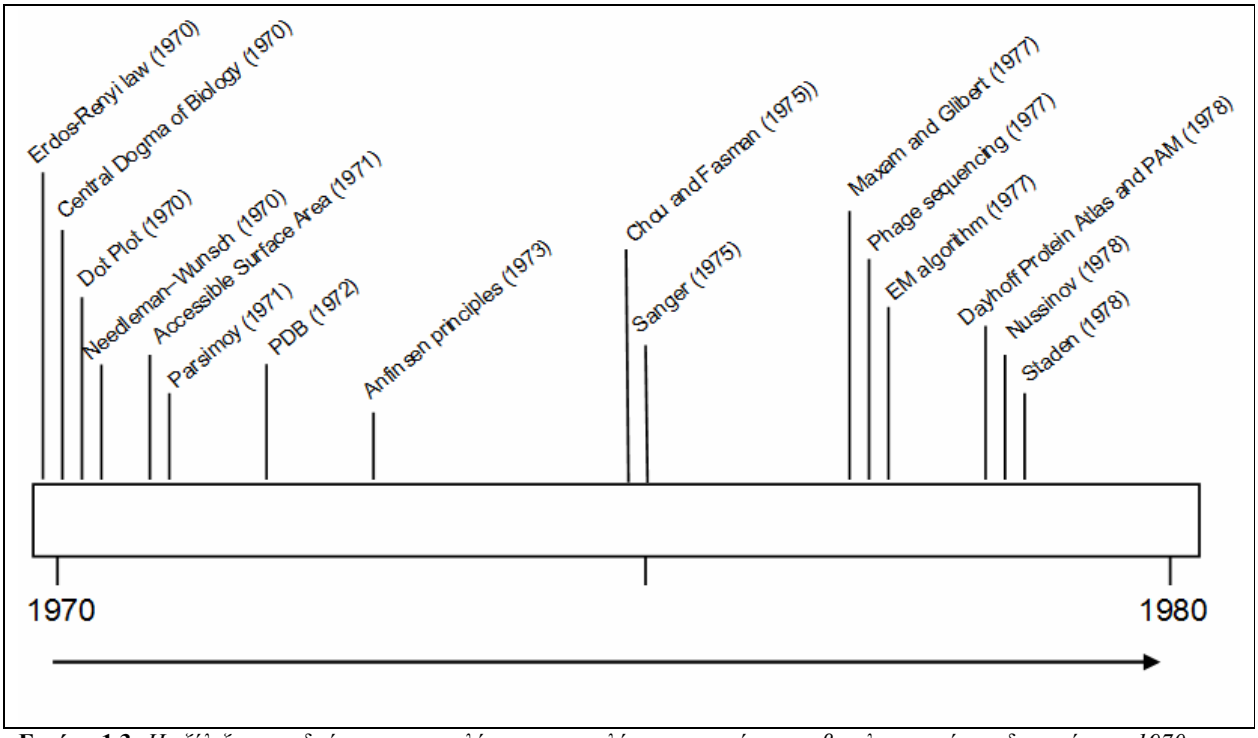
1.1.2. Η δεκαετία του 1970

Την επόμενη δεκαετία, η έρευνα συνεχίστηκε με αυξανόμενο ρυθμό (Εικόνα 1.3). Μια από τις πιο σημαντικές συνεισφορές αυτής της περιόδου, με ευρύτερες συνέπειες για τις βιοεπιστήμες, ήταν η σύγκλιση της κλασικής πληθυσμιακής γενετικής με τη μοριακή εξέλιξη, με αφορμή τις εργασίες του Kimura που είδαμε πριν. Έτσι, φτάσαμε στην εμφάνιση της θεωρίας της ουδέτερης εξέλιξης και στην υπόθεση του σταθερού ρυθμού εξελικτικών αλλαγών, η οποία είναι γνωστή ως το «μοριακό ρολόι». Την ίδια εποχή εμφανίστηκε και ο γνωστός αλγόριθμος του Fitch για την φειδωλή ανακατασκευή φυλογενετικών δέντρων με τη χρήση αλληλουχιών (μέθοδος της μέγιστης φειδωλότητας). Καθώς ο γενετικός κώδικας είχε αποκαλυφθεί, και είχε διαλευκανθεί ο ρόλος των RNA στην μεταγραφή και τη μετάφραση, το κεντρικό δόγμα της βιολογίας διατυπώθηκε από τον Crick το 1970. Την ίδια δεκαετία, εμφανίστηκαν και οι πρώτες μεθοδολογίες αλληλούχισης νουκλεϊκών οξέων από τους Sanger και Maxam-Glbert, μεθοδολογίες που έδωσαν ώθηση στη μελέτη των γονιδιωμάτων και με διάφορες παραλλαγές και τροποποιήσεις έχουν φτάσει μέχρι σήμερα, στις σύγχρονες μεθόδους αλληλούχισης.

Στο επίπεδο των πρωτεϊνών, οι πρωτοποριακές εργασίες του Anfinsen για τις αρχές που καθορίζουν την πρωτεϊνική δομή έδωσαν ώθηση στην έρευνα σε αυτό το πεδίο. Τότε εμφανίστηκαν οι πρώτες μέθοδοι υπολογισμού της προσβασιμότητας στο διαλύτη, όσο και οι πρώτες εργασίες για τις προτιμήσεις των αμινοξέων για τα διάφορα στοιχεία δευτεροταγούς δομής, οι οποίες οδήγησαν στον πρώτο αλγόριθμό πρόγνωσης της δευτεροταγούς δομής πρωτεϊνών από τους Chou και Fasman το 1975 (ενώ φυσικά ακολούθησαν και άλλοι τα επόμενα χρόνια). Επίσης λίγο αργότερα, εμφανίστηκαν και οι πρώτες προσπάθειες πρόγνωσης της δομής των RNA. Τα μοριακά γραφικά, αλλά και οι πρώτες προσπάθειες προσομοίωσης του πρωτεϊνικού διπλώματος με μοριακή δυναμική, εμφανίστηκαν επίσης εκείνη την εποχή. Παρόμοια με τις πρωτεΐνες, εμφανίστηκαν και οι πρώτοι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής του RNA με τις πρωτοποριακές εργασίες της Nussinov.

Μια από τις πιο σημαντικές αλγοριθμικές συνεισφορές στην υπολογιστική βιολογία που συνέβησαν τη δεκαετία του 1970, ήταν η εμφάνιση των αλγορίθμων δυναμικού προγραμματισμού για τη στοίχιση βιολογικών αλληλουχιών (κυρίως πρωτεϊνών), με πρώτο τον αλγόριθμο για ολική στοίχιση των Needleman και Wunsch το 1970, ενώ ακολούθησαν και άλλες προσεγγίσεις και μελέτες στη μεθοδολογία και τα στατιστικά της στοίχισης, ενώ το 1970 έκανε και την εμφάνισή του το dot-plot. Τέλος, αυτή τη δεκαετία εμφανίστηκαν και οι πρώτες βάσεις βιολογικών δεδομένων. Η PDB εμφανίστηκε το 1972 (όταν υπήρχαν μόλις 10 τρισδιάστατες δομές πρωτεϊνών), ενώ η Dayhoff παρουσίασε το 1978 και την πρώτη συλλογή πρωτεϊνικών αλληλουχιών οι οποίες ήταν γνωστές εκείνα τα χρόνια, μια συλλογή που κατά κάποιον τρόπο

μπορεί να θεωρηθεί ο πρόδρομος της PIR. Τέλος, τα πρώτα προγράμματα Η/Υ για απλές αναλύσεις σε βιολογικές αλληλουχίες έκαναν την εμφάνισή τους (μετάφραση μιας κωδικής αλληλουχίας, εύρεση προτύπων, αναγνώριση υποκινητών και θέσεων δράσης περιοριστικών ενζύμων κ.ο.κ.). Όπως είναι φανερό, ήδη από τη δεκαετία του 1970 είχε ήδη σχηματιστεί μια καθαρή εικόνα του ερευνητικού πεδίου της Βιοπληροφορικής. Υπήρχαν οι αλγόριθμοι στοίχισης, η θεωρία της μοριακής εξέλιξης και η ποσοτικοποίηση των εξελικτικών αλλαγών, η κατασκευή φυλογενετικών δέντρων, οι μεθοδολογίες μελέτης και πρόγνωσης της δευτεροταγούς και τριτοταγούς δομής των πρωτεϊνών και οι πρώτες βιολογικές βάσεις δεδομένων.



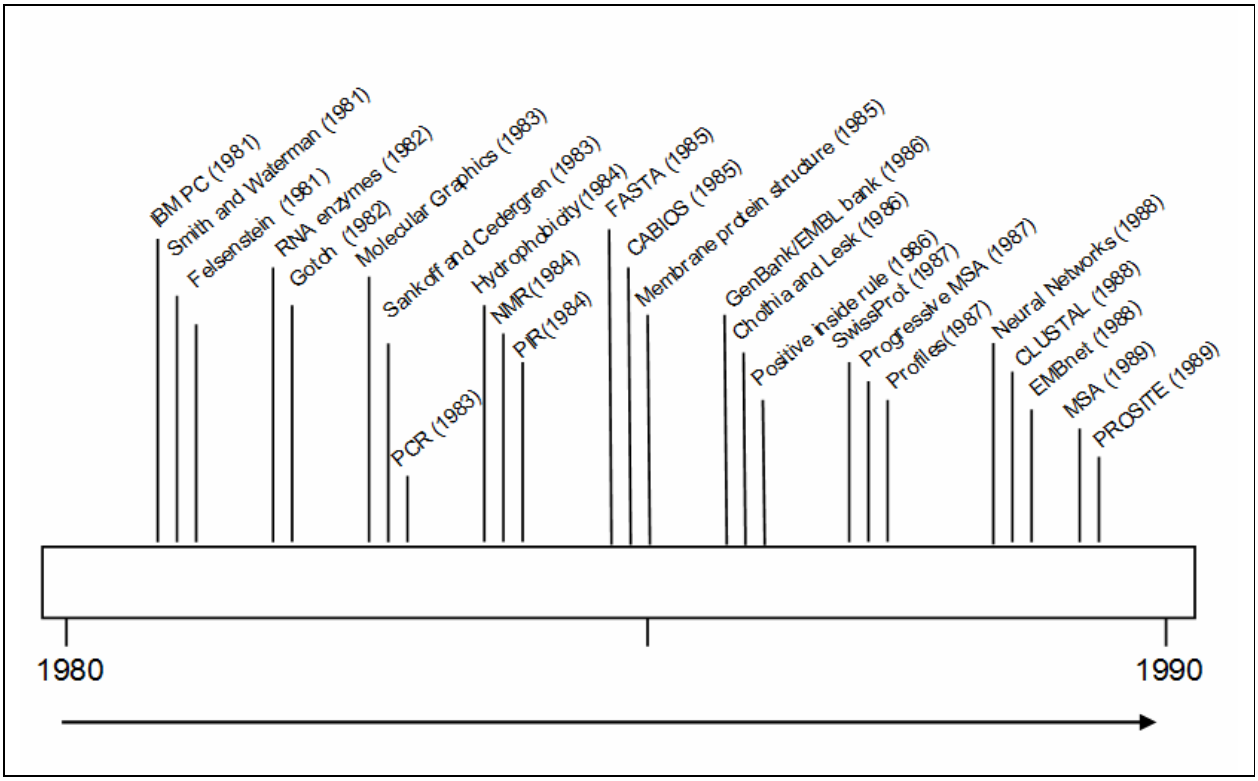
Εικόνα 1.3: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληφορική στη δεκαετία του 1970*

1.1.3. Η δεκαετία του 1980

Η δεκαετία του 1980 ήταν η δεκαετία στην οποία το πεδίο της υπολογιστικής βιολογίας πήρε πλέον μια ξεκάθαρη μορφή, σαν ένας ξεχωριστός κλάδος θέτοντας τα δικά του προβλήματα αλλά και παρουσιάζοντας και τα σημαντικά του επιτεύγματα. Αρχίζουν να κάνουν μαζική εμφάνιση οι αντίστοιχες δημοσιεύσεις στα υψηλού κύρους βιολογικά περιοδικά (Science, Nature, Nucleic Acid Research), ενώ και τα πρώτα εξειδικευμένα περιοδικά κάνουν την εμφάνισή τους (Computer Applications in Biosciences). Φυσικά, πρέπει να έχουμε στο μυαλό μας ότι την εποχή αυτή είχε αρχίσει να γίνεται διαδεδομένη η χρήση υπολογιστικών συστημάτων και έτσι πολλές από τις παρακάτω ανακαλύψεις ακολούθησαν και επωφελήθηκαν από την πρόοδο στον τομέα του υλικού και του λογισμικού.

Στο πεδίο της ανάλυσης αλληλουχιών μακρομορίων, η μελέτη πάνω στους αλγόριθμους στοίχισης και στις αποτελεσματικές υλοποιήσεις τους συνεχίστηκε με εντατικό ρυθμό. Βασικό ρόλο έπαιξαν σε αυτή την πρόοδο η ανακάλυψη του αλγορίθμου τοπικής στοίχισης από τους Smith και Waterman το 1981, οι αλγόριθμοι προσεγγιστικού ταιριάσματος συμβολοσειρών, η μελέτη των στατιστικών ιδιοτήτων της στοίχισης από τους Aratia, Waterman και Karlin, αλλά και οι πρώτες αποτελεσματικές υλοποιήσεις για γρήγορη στοίχιση και αναζήτηση ομοιότητας σε μια βάση δεδομένων (FASTA). Παράλληλα έγιναν οι πρώτες θεωρητικές επεξεργασίες της πολλαπλής στοίχισης, επινοήθηκε η ιεραρχική πολλαπλή στοίχιση και παρουσιάστηκε το CLUSTAL. Ιδιαίτερα σημαντική επινόηση αυτής της περιόδου ήταν τα προφίλ αλληλουχιών (sequence profiles) τα οποία αποτέλεσαν πανίσχυρο εργαλείο στη μελέτη των πρωτεϊνικών οικογενειών, εφαρμόστηκαν σε πάρα πολλά παραδείγματα με εντυπωσιακά αποτελέσματα και εξακολουθούν να χρησιμοποιούνται μέχρι σήμερα. Τέλος, πρέπει να σημειώσουμε ότι την εποχή αυτή εμφανίστηκαν και τα πρώτα βιβλία σχετικά με την υπολογιστική ανάλυση αλληλουχιών.

Η πρόοδος στις μεθόδους αλληλούχισης DNA, η εμφάνιση της PCR, αλλά και η ραγδαία βελτίωση στις τεχνικές προσδιορισμού της τρισδιάστατης δομής των μακρομορίων, οδήγησαν την εποχή αυτή στη ραγδαία αύξηση του όγκου των δεδομένων και στη δημιουργία μεγαλύτερων και πιο οργανωμένων βάσεων βιολογικών δεδομένων (φυσικά, και αυτή η δραστηριότητα αναπτύχθηκε παράλληλα με τις εξελίξεις στα πληροφοριακά συστήματα και τα συστήματα βάσεων δεδομένων). Έτσι, το 1986 έκαναν την εμφάνιση τους οι δύο πιο γνωστές μέχρι σήμερα βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών (GenBank και EMBL Data Library), ενώ η SwissProt, η βάση δεδομένων των πρωτεϊνικών αλληλουχιών εμφανίστηκε το 1987. Την ίδια εποχή έκαναν την εμφάνισή τους προτάσεις για δημιουργία δικτύων που θα διευκόλυναν την υπολογιστική έρευνα στη βιολογία (EMBnet και BIONET), ενώ εμφανίστηκαν και οι πρώτοι κατάλογοι με σχετικό λογισμικό (LiMB). Τέλος, οι ερευνητικοί οργανισμοί όπως το NIH και το EMBL άρχισαν τη δημιουργία εξειδικευμένων τμημάτων αφιερωμένων στην έρευνα στην υπολογιστική βιολογία.



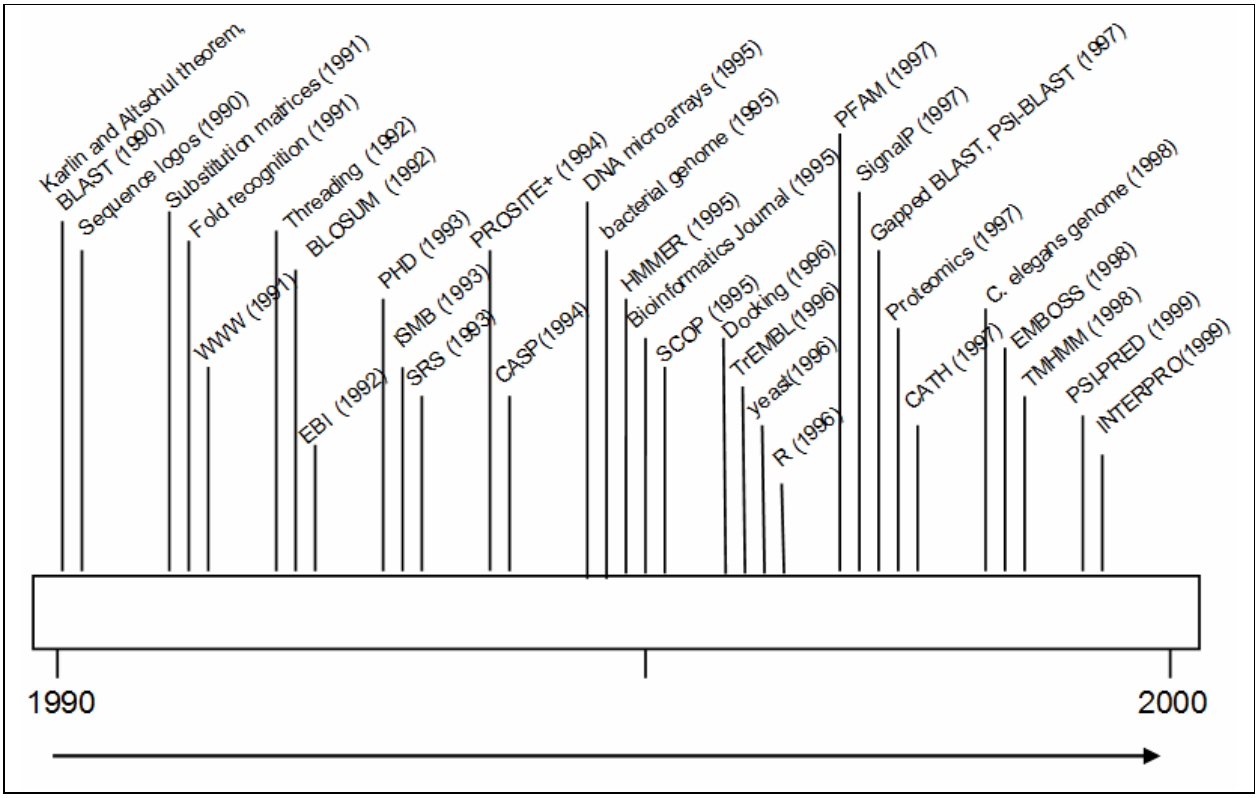
Εικόνα 1.4: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληφορική στη δεκαετία του 1980*

Στον τομέα της ανάλυσης και πρόγνωσης της δομής των πρωτεϊνών, έγιναν επίσης σημαντικές εξελίξεις. Η κρυσταλλογραφία συνέχισε να βελτιώνεται, και εμφανίστηκε και το NMR ενώ οι μεθοδολογίες αναπαράστασης μοριακών τρισδιάστατων δομών εξακολούθησαν να εξελίσσονται παράλληλα με τις εξελίξεις στον τομέα των γραφικών και της υπολογιστικής γεωμετρίας. Με την αύξηση των δομών, αλλά και την εξάπλωση των αλγόριθμων στοίχισης, εμφανίστηκαν και οι πρώτες προσπάθειες αυτόματης προτυποποίησης πρωτεϊνικών δομών με βάση την ομολογία (homology modelling). Η αύξηση των πρωτεϊνικών δομών και η κατάταξή τους σε οικογένειες και διπλώματα (δομικά μοτίβα), οδήγησε και στις πρώτες προσπάθειες μελέτης των κατηγοριών πρωτεϊνικού διπλώματος και προσπάθειες πρόγνωσής του. Επιπλέον, οι μελέτες των δομών κατέληξαν στο πολύ σημαντικό συμπέρασμα ότι οι δομές των πρωτεϊνών συντηρούνται περισσότερο από ότι οι αλληλουχίες τους. Οι αλγόριθμοι πρόγνωσης δευτεροταγούς δομής συνέχισαν να εξελίσσονται, δεχόμενες και τη βοήθεια νέων εξελίξεων στην τεχνητή νοημοσύνη (νευρωνικά δίκτυα), ενώ οι πρώτες κρυσταλλικές δομές μεμβρανικών πρωτεϊνών έδωσαν το έναυσμα για την ανάπτυξη των πρώτων μεθόδων ανάλυσης της αλληλουχίας των πρωτεϊνών αυτών (διαγράμματα υδροφοβικότητας, υδροφοβικές ροπές, positive inside rule) και οι πρώτοι αλγόριθμοι πρόγνωσης έκαναν την εμφάνισή τους. Παρόμοια με τις πρωτεΐνες, εξαπλώθηκαν και οι αλγόριθμοι πρόγνωσης της δευτεροταγούς δομής του RNA.

Στον τομέα της φυλογενετικής ανάλυσης, την εποχή αυτή προτάθηκε ο αλγόριθμος του Fenselstein για την εκτίμηση φυλογενετικών δέντρων μέσω της μέγιστης πιθανοφάνειας, μια πολύ σημαντική ανακάλυψη που έδωσε ώθηση στο αντίστοιχο πεδίο (ενώ παράλληλα αναπτύχθηκαν και πολλά από τα γνωστά μέχρι σήμερα μαθηματικά μοντέλα για την αντικατάσταση βάσεων σε φυλογενετικές μελέτες). Την ίδια εποχή έγιναν σημαντικές συνεισφορές και στη μελέτη των εξελικτικών σχέσεων των πρωτεϊνών (μιλήσαμε ήδη για την ανακάλυψη ότι οι δομές συντηρούνται περισσότερο από την αλληλουχία). Μελετήθηκαν οι στατιστικές ιδιότητες των πινάκων αντικατάστασης αμινοξέων (PAM), μελετήθηκε έντονα το φαινόμενο της ομολογίας, αλλά και περιπτώσεις ομοιότητας λόγω συγκλίνουσας εξέλιξης, ενώ έγιναν πολλές μελέτες της εξελικτικής ιστορίας συγκεκριμένων πρωτεϊνικών οικογενειών οι οποίες είχαν ευρύτερη σημασία στη βιολογία (π.χ. ανοσοσφαιρίνες, πρωτεάσες, κυτοχρώματα, ριβονουκλεάσες κ.ο.κ.). Τέλος, έγιναν σημαντικά βήματα στην εξελικτική μελέτη των γονιδιωμάτων καθώς μελετήθηκαν οι φυλογενετικοί δείκτες όπως το rRNA, αλλά και η εξελικτική ιστορία των εσωνίων, των εξωνίων και της συρραφής.

1.1.4. Η δεκαετία του 1990

Η δεκαετία του 1990 ήταν η δεκαετία κατά την οποία η έρευνα στην υπολογιστική βιολογία εκτινάχθηκε (θα δούμε παρακάτω και εμπειρικά μετρήσιμα δεδομένα για αυτό). Φυσικά, για άλλη μια φορά δεν πρέπει να αμελήσουμε να αναφέρουμε ότι η δεκαετία αυτή σημαδεύτηκε επίσης από την ανάπτυξη του διαδικτύου και του παγκοσμίου ιστού, αλλά και από την εξάπλωση των προσωπικών Η/Υ. Πρέπει να σημειώσουμε επίσης, ότι η ευρεία χρήση του όρου «βιοπληροφορική» συντελέστηκε μέσα στη δεκαετία του 1990. Ενδεικτικά, το πιο γνωστό περιοδικό του χώρου, το Bioinformatics, πήρε το όνομα αυτό το 1995 αλλάζοντας το προηγούμενο όνομα «Computer Applications in the Biosciences» (CABIOS).



Εικόνα 1.5: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληφορική στη δεκαετία του 1990*

Η δεκαετία αυτή σημαδεύτηκε από μια σειρά μεγάλες ανακαλύψεις. Στον τομέα της στοίχισης αλληλουχιών, η πιο σημαντική ίσως δημοσίευση όλων των εποχών στο χώρο, αφορά το BLAST (Basic Local Alignment Search Tool), από επιστήμονες του NCBI το 1990. Το BLAST «πάτησε» πάνω στις ανακαλύψεις για τη στατιστική κατανομή του score της τοπικής στοίχισης (το γνωστό θεώρημα Karlin-Altschul) και πραγματικά ήταν μια επαναστατική συμβολή στον τρόπο που θα διεξάγεται από κει και πέρα η αναζήτηση

ομοιότητας σε βάσεις δεδομένων και η στοίχιση, καθώς ήταν πιο γρήγορο από κάθε άλλο αλγόριθμο επιτρέποντας ταχείες αναζητήσεις, αλλά έδινε και για πρώτη φορά μια εκτίμηση για τη στατιστική σημαντικότητα των στοιχίσεων. Στην πρώτη του έκδοση δεν παρήγαγε στοιχίσεις με κενά, αλλά στη δεύτερη, παρείχε και αυτή τη δυνατότητα, ενώ περιλάμβανε και άλλες παραλλαγές όπως το PSI-BLAST. Επιπλέον, τα προγράμματα πολλαπλής στοίχισης έκαναν τη δυναμική τους εμφάνιση (CLUSTAL) με εκδόσεις για μαζική χρήση σε Η/Υ, δίνοντας ακόμα και εκδόσεις για παραθυρικό περιβάλλον.

Στον τομέα της ανάλυσης των πρωτεϊνικών δομών και της πρόγνωσης έγιναν επίσης μεγάλες ανακαλύψεις. Εμφανίστηκαν τα πρώτα προγράμματα ευρείας χρήσης για την οπτικοποίηση και την ανάλυση πρωτεϊνικών δομών, όπως το Rasmol και το Kinemage, ενώ έγιναν και οι πρώτες επιτυχημένες προσπάθειες για ύφανση πρωτεϊνών (threading), αλλά και για αγκυροβόληση (docking) πρωτεϊνικών δομών. Στον τομέα της πρόγνωσης της δευτεροταγούς δομής, η χρήση νευρωνικών δικτύων παράλληλα με τη χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων, έδωσε για πρώτη φορά ποσοστά επιτυχίας πάνω από 70% και άνοιξε ολόκληρες λεωφόρους στη μελέτη των αλγορίθμων πρόγνωσης με εφαρμογές και σε πλήθος άλλων περιπτώσεων (διαμεμβρανικές πρωτεϊνες, προσβασιμότητα του διαλύτη, κ.ο.κ.). Παράλληλα, ξεκίνησαν και οι διαγωνισμοί/συνέδρια του CASP.

Οι πρώτες επιτυχημένες προσπάθειες αλληλούχισης ολόκληρων γονιδιωμάτων, πρώτα βακτηρίων και στη συνέχεια και ευκαρυωτικών οργανισμών, άνοιξαν επίσης νέους δρόμους στη συγκριτική γονιδιωματική, ενώ πυροδότησαν και την ανάπτυξη των πρώτων αλγορίθμων εύρεσης γονιδίων (gene finders). Η δεκαετία αυτή, σηματοδότησε επίσης την εμφάνιση των μικροσυστοιχιών DNA για τη μέτρηση της γονιδιακής έκφρασης, τεχνολογία που είχε, όπως θα δούμε στη συνέχεια, μεγάλη επίδραση τόσο στη Βιοπληροφορική όσο και στην Ιατρική Πληροφορική και τη Βιοστατιστική, και σηματοδότησε την απαρχή της λειτουργικής γονιδιωματικής.

Στον τομέα των βάσεων δεδομένων, η εκθετική αύξηση των δεδομένων όλων των κατηγοριών συνεχίστηκε και μια σειρά νέες βάσεις δεδομένων αναπτύχθηκαν. Ανάμεσά τους ήταν βάσεις με δομικές ταξινομήσεις των πρωτεϊνών (όπως η SCOP και η CATH), αλλά και βάσεις με ταξινομήσεις βασισμένες σε χαρακτηριστικά πρότυπα (patterns) της ακολουθίας, όπως η PROSITE, η PFAM και τελικά η INTERPRO. Επίσης, μια πολύ σημαντική εξέλιξη αυτής της περιόδου, ήταν η ίδρυση του EBI (European Bioinformatics Institure), του μεγαλύτερου ινστιτούτου βιοπληροφορικής της Ευρώπης, το οποίο ιδρύθηκε στη Μεγάλη Βρετανία (Hinxton) το 1992 μέσα από μια κοινοπραξία του EMBL και του Welcome Trust. Στο EBI στεγάστηκαν αρχικά οι βάσεις δεδομένων του EMBL, EMBL-Bank και SwissProt-TrEMBL και δημιουργήθηκαν ερευνητικές ομάδες για να συνδράμουν στα διάφορα γονιδιωματικά προγράμματα εκείνης της εποχής, ενώ λίγο αργότερα λειτούργησε και η TrEMBL. Τέλος, το 1993 ξεκίνησαν τα συνέδρια ISMB και λίγα χρόνια αργότερα ιδρύθηκε η ISCB.

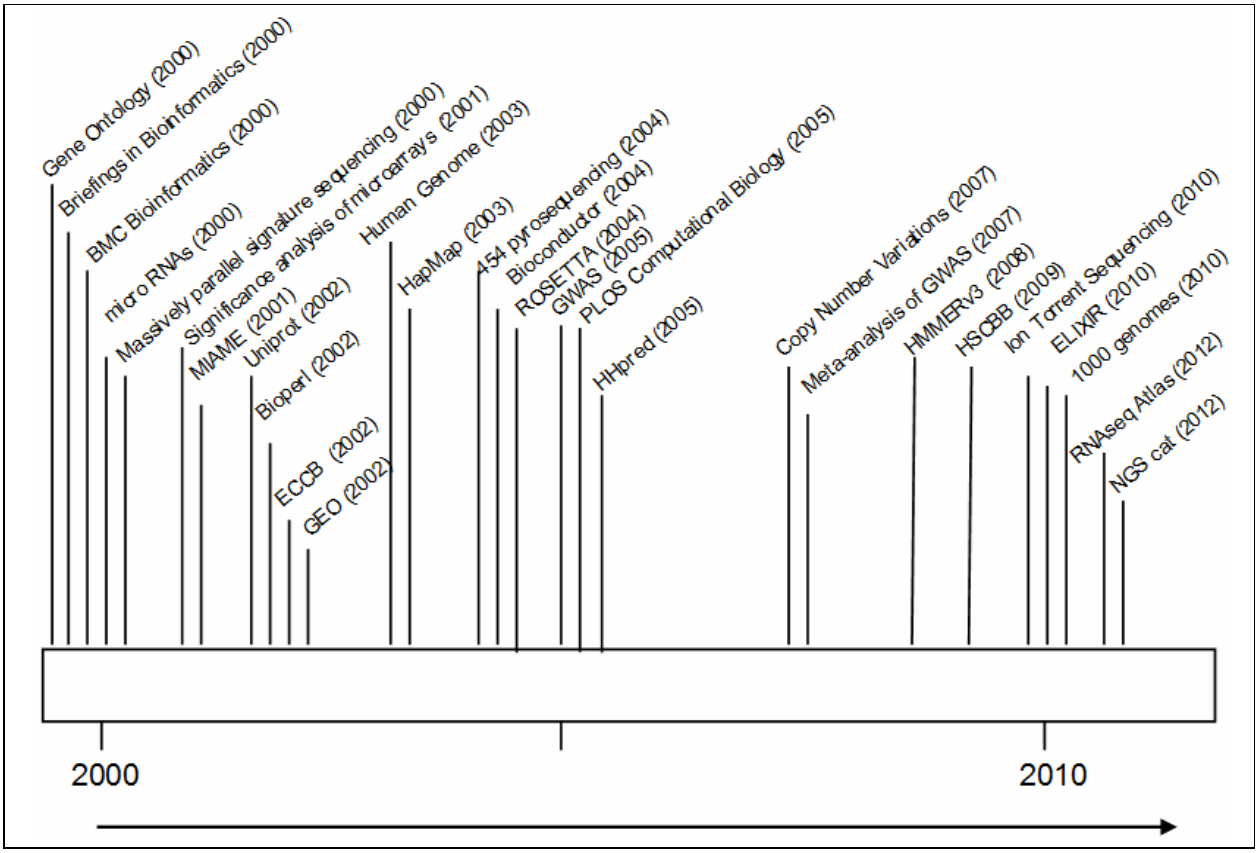
Τέλος, τη δεκαετία αυτή έκανε την εμφάνισή της, μετά τις πρωτοποριακές εργασίες των Krogh, Eddy, Hughey κλπ, και μια μεθοδολογία που θα επικρατούσε τα επόμενα χρόνια στην ανάλυση αλληλουχιών, το Hidden Markov Model (HMM), το οποίο βρήκε εφαρμογές τόσο στην μοντελοποίηση των πολλαπλών στοιχίσεων και την αναζήτηση μακρινών ομοιοτήτων, όσο και στις μεθόδους πρόγνωσης. Το γνωστό πακέτο HMMER για πολλαπλές στοιχίσεις και αναζήτησεις μακρινών ομολόγων με profile HMM, πάνω στο οποίο βασίζεται η βάση δεδομένων πρωτεϊνικών οικογενειών PFAM, έκανε την εμφάνιση του εκείνη την περίοδο, ενώ το ίδιο συνέβη και για δυο από τους πιο επιτυχημένους αλγόριθμους πρόγνωσης, το TMHMM για τις μεμβρανικές πρωτεϊνες, και το SignalP για τις σηματοδοτικές αλληλουχίες. Το HMM αξίζει μια ειδική αναφορά, γιατί παρόλο που σαν μαθηματική μέθοδος ήταν γνωστή από καιρό και είχε χρησιμοποιηθεί στην αναγνώριση ομιλίας, η υιοθέτησή του σε μεθόδους υπολογιστικής βιολογίας, έδωσε νέα πνοή και στην ίδια την αναζήτηση μεθοδολογίας καθώς μια σειρά από αλγόριθμους και τροποποιήσεις του μοντέλου εμφανίστηκαν ειδικά για τα προβλήματα της βιολογίας (το profile HMM, οι αλγόριθμοι για σημασμένες αλληλουχίες, αλλά και μια σειρά αλγόριθμοι εκπαίδευσης και αποκωδικοποίησης). Σήμερα, δεν νοείται κείμενο, ακόμα και εισαγωγικό, στη βιοπληροφορική που να μην περιγράφει το HMM.

1.1.5. Η «σύγχρονη» εποχή

Τέλος, η εποχή μετά το 2000 σηματοδότησε την ώριμη φάση της υπολογιστικής βιολογίας, καθώς η διδασκαλία της έχει γίνει βασική πλέον και σε προπτυχιακό αλλά και σε μεταπτυχιακό επίπεδο, ιδρύονται και επεκτείνονται επαγγελματικές οργανώσεις, κυκλοφορούν όλο και περισσότερα εξειδικευμένα επιστημονικά περιοδικά κ.ο.κ. Η ολοκλήρωση του προγράμματος προσδιορισμού του ανθρώπινου γονιδιώματος, μπορεί να

μη δικαίωσε όλες τις αρχικές προσδοκίες («*θα βρούμε το φάρμακο για κάθε ασθένεια*») αλλά σίγουρα άνοιξε νέους δρόμους σε μια σειρά από κλάδους. Για παράδειγμα, ο μεγάλος ρυθμός προσδιορισμού των γονιδιωμάτων, οδήγησε στην αλματώδη ανάπτυξη της γονιδιωματικής στις διάφορες μορφές της. Συγκριτική γονδιωματική για τη σύγκριση γονιδιωμάτων, λειτουργική γονιδιωματική για τις μελέτες γονιδιακής έκφρασης με μικροσυστοιχίες και δομική γονιδιωματική για τη μαζική παραγωγή πρωτεϊνών για δομικές μελέτες και κρυσταλλογραφία ακτίνων Χ. Παράλληλα, εμφανίστηκαν οι τεχνικές αλληλούχισης νέας γενιάς (Next Generation Sequencing), οι οποίες έδωσαν νέα πνοή σε μελέτες γονιδιακής έκφρασης (RNAseq), έκαναν εύκολο τον εντοπισμό πολυμορφικών θέσεων και αναμένεται να επηρεάσουν και την προσωποποιημένη ιατρική. Με όλα αυτά τα δεδομένα. δημιουργήθηκε μια μεγάλη ανάγκη για περισσότερους αλγόριθμους πρόγνωσης έτσι ώστε να τεθεί 'σε τάξη' ο τεράστιος αυτό όγκος δεδομένων, αλλά και μια μεγάλη ανάγκη για δημιουργία εξειδικευμένων βάσεων δεδομένων και οντολογιών που να τις περιγράφουν. Την εποχή αυτή, είδαμε την έκρηξη των διαδικτυακών εφαρμογών (web-servers), αλλά και του ελεύθερου λογισμικού βιοπληροφορικής. Επιπλέον, τα προγράμματα μετα-γονδιωματικής (meta-genomics) έθεσαν νέα αλγοριθμικά προβλήματα στην εύρεση γονιδίων και την ομαδοποίηση δεδομένων.

Ειδικά στις βάσεις δεδομένων, μέσα στη δεκαετία του 2000, εκτός από την ανάπτυξη μικρών εξειδικευμένων βάσεων δεδομένων, είδαμε και τις πρώτες συγχωνεύσεις των μεγάλων βάσεων δεδομένων, καθώς η SwissProt και η PIR ένωσαν τις προσπάθειές τους σε μια προσπάθεια να ανταπεξέλθουν στον τεράστιο όγκο δεδομένων, σχηματίζοντας την Uniprot (η οποία πλέον στεγάζεται στο EBI). Οι προβλέψεις για το μέλλον είναι κάπως δυσοίωνες, καθώς με τη συνεχόμενη εκθετική αύξηση των δεδομένων, σε λίγα χρόνια θα υπάρχει μεγάλο πρόβλημα αποθήκευσης και διαμοιρασμού των δεδομένων, γι' αυτό και έχει ξεκινήσει από το EBI η πρωτοβουλία του ELIXIR να διανεμηθούν, κατά κάποιον τρόπο, οι δημόσια διαθέσιμες βάσεις σε διάφορες χώρες και φορείς.



Εικόνα 1.6: *Η εξέλιξη των ιδεών με τις μεγαλύτερες ανακαλύψεις σχετικές με τη βιοπληφορική στη δεκαετία του 2000*

Η γνώση της αλληλουχίας του ανθρώπινου γονιδιώματος έδωσε επίσης μεγάλη ώθηση στη Γενετική Επιδημιολογία, καθώς με τον εντοπισμό εκατομμυρίων πολυμορφισμών (SNPs) και τη χρήση της τεχνολογίας των GWAS (Genome-Wide Association Studies), μας δόθηκε η δυνατότητα να κάνουμε μαζικά μελέτες

γενετικής συσχέτισης, μελετώντας ταυτόχρονα εκατομμύρια πολυμορφισμούς σε μαζική κλίμακα, όπως ακριβώς και με τις μικροσυστοιχίες DNA. Παράλληλα, έγιναν μελέτες για την απλοτυπική σύσταση και την προέλευση των ανθρώπινων πληθυσμών, την κατανομή τους, το βαθμό ανασυνδυασμού κλπ (HapMap project), ενώ μεγάλη αποδοχή έχουν λάβει οι μεθοδολογίες μετα-ανάλυσης και ενοποίησης δεδομένων (data integration). Οι περιοχές αυτές, είναι περιοχές που πλέον η Βιοπληροφορική έρχεται σε επαφή με την Γενετική και την Επιδημιολογία και τη Βιοστατιστική. Επιπλέον δε, γνώση της αλληλουχίας του γονιδιώματος, επηρρέασε και άλλους κλάδους όπως την Πρωτεομική (Proteomics), ενώ η λεπτομερής μελέτη του, έδωσε και νέες βιολογικές ανακαλύψεις, όπως τα μη κωδικά RNA (ncRNA), για τα οποία αναπτύχθηκαν μια σειρά αλγόριθμων και μεθοδολογιών για την πρόγνωση και τη μελέτη τους, αλλά και των μεγάλων επαναληπτικών αλληλουχιών (Copy Number Variations- CNVs). Και οι δύο περιπτώσεις, δεν ήταν προηγουμένως γνωστές και πλέον βρίσκονται στο επίκεντρο της μοριακής έρευνας.

Αλγοριθμικά, εμφανίστηκε δυναμικά η χρήση των Support Vector Machines (SVMs) σε προβλήματα πρόγνωσης, μια μεθοδολογία που αποδείχθηκε πιο αποτελεσματική από τα νευρωνικά δίκτυα σε κάποιες περιπτώσεις. Επίσης, παρουσιάστηκαν ολοκληρωμένοι αλγόριθμοι για σύγκριση HMM-HMM, αλλά και η νέα έκδοση του HMMER η οποία στηρίζεται σε μια σειρά θεωρητικές βελτιώσεις που βελτιώνουν δραματικά την ταχύτητά του και φιλοδοξούν πλέον να το καταστήσουν αντικαταστάτη του BLAST. Επίσης, έγιναν μεγάλες προόδοι στην ab initio πρόγνωση δομής πρωτεϊνών. Τέλος, τη δεκαετία αυτή, ακολουθώντας την τεράστια αύξηση των βάσεων δεδομένων και των οντολογιών, έκανε την εμφάνιση της η βιολογία συστημάτων, η οποία μελετάει πλέον πολύπλοκα δίκτυα με τις αλληλεπιδράσεις των μερών τους, αντι για μεμονωμένες οντότητες, αλλά και οι οντολογίες. Τέτοια δίκτυα είναι τα δίκτυα πρωτεϊνικών αλληλεπιδράσεων, τα ρυθμιστικά δίκτυα, τα τροφικά δίκτυα κλπ. Στην ανάπτυξη αυτή, εκτός από την ποσότητα των δεδομένων και την αυξημένη διαθέσιμη υπολογιστική ισχύ, σημαντικό ρόλο έπαιξαν και οι εξελίξεις στη μαθηματική θεωρία των γράφων και στη θεωρητική πληροφορική.

Στα επόμενα χρόνια, αναμένεται ο ρόλος και η μορφή της Υπολογιστικής Βιολογίας να αλλάξει όλο και περισσότερο, ακολουθώντας τις ραγδαίες εξελίξεις της τεχνολογίας και την ολοένα μεγαλύτερη συσώρευση μοριακών δεδομένων. Δεν μπορούμε να προβλέψουμε ακριβώς ποια θα είναι η μορφή αυτή, αλλά σίγουρα οι εξελίξεις στην προσωποιημένη ιατρική, στην αλληλούχιση, στη νανοτεχνολογία, στο βιολογικό υπολογισμό αλλά και στην ίδια την επιστήμη υπολογιστών, θα επηρεάσουν και τον τρόπο που η Υπολογιστική Βιολογία προσεγγίζει τα πράγματα και εξερευνά νέα μονοπάτια (C. A. Ouzounis, 2012).

1.2. Η διεπιστημονικότητα της βιοπληροφορικής

Όπως έγινε ελπίζω κατανοητό από τα προηγούμενα η Βιοπληροφορική ή, καλύτερα, η Υπολογιστική Βιολογία είναι διεπιστημονικός κλάδος που έλκει την καταγωγή του από τη μοριακή βιολογία και ιδιαίτερα από τη μελέτη των βιολογικών αλληλουχιών και των δομών. Ο ορισμός που δώσαμε, είναι περισσότερο συμβατός με τον ορισμό που δίνει το NCBI (http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html):

«Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information»

ή, τον αντίστοιχο ορισμό του Luscombe (Luscombe, Greenbaum, & Gerstein, 2001):

«Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale»

και αυτόν του Fredj Tekaia:

«The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information».

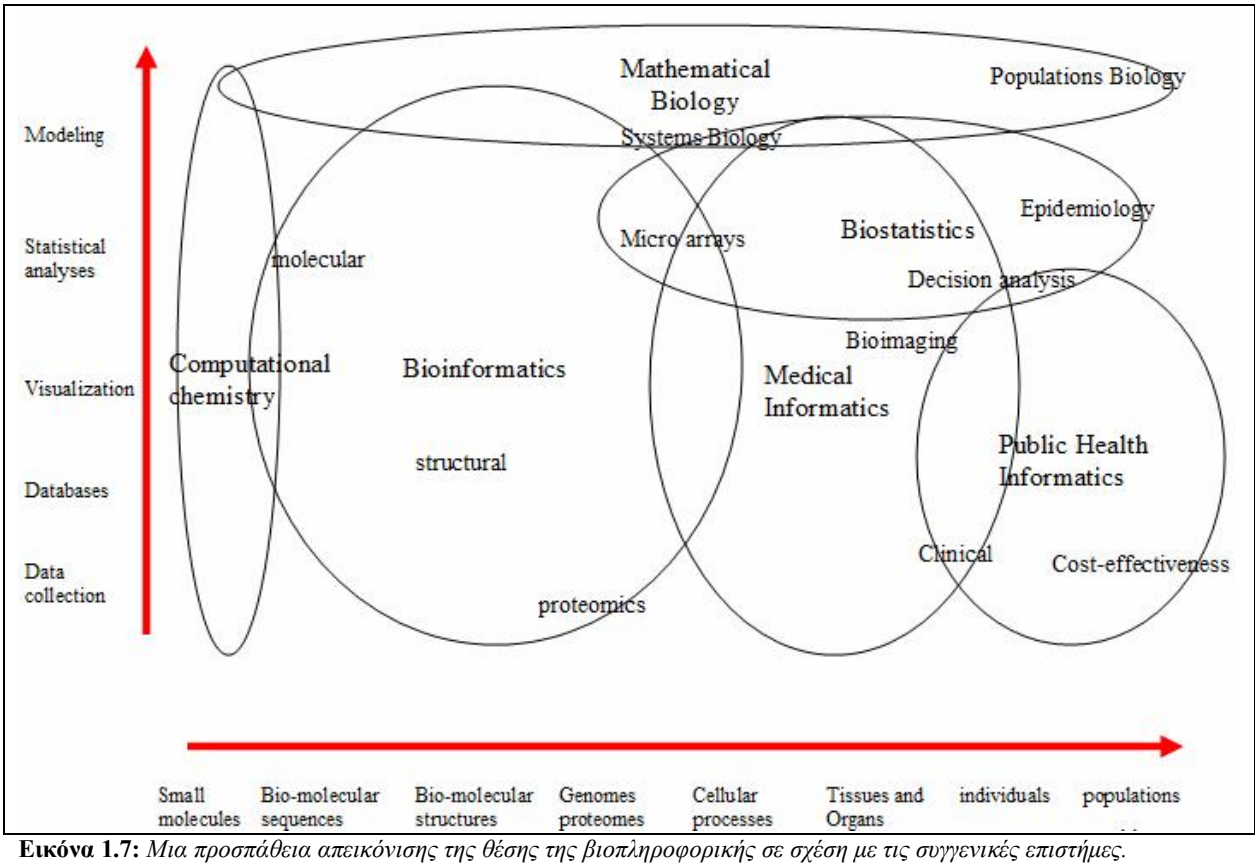
αλλά υπάρχουν και ορισμοί που διαφωνούν ριζικά, ιδιαίτερα ορισμοί που κάνουν σαφή διάκριση μεταξύ της βιοπληροφορικής (διαχείριση μεγάλου όγκου δεδομένων και βάσεων δεδομένων) και της Υπολογιστικής Βιολογίας (ανάπτυξη αλγόριθμων και μεθοδολογιών). Για παράδειγμα ο Richard Durbin λέει ότι:

«I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information».

Όπως και να έχει, η International Society for Computational Biology (ISCB) αποφεύγει τους ορισμούς, και αυτοπροσδιορίζεται, κάπως πιο γενικά, ως:

«a scholarly society dedicated to advancing the scientific understanding of living systems through computation».

Μερικοί, πάνε αυτόν τον ορισμό ακόμα πιο μακριά και δέχονται ότι και δραστηριότητες που σήμερα ταξινομούνται στην Ιατρική Πληροφορική και τη Βιοϊατρική Τεχνολογία, όπως η ανάλυση ιατρικών εικόνων και η διαχείριση του ιατρικού φακέλου του ασθενούς, ανήκουν στη Βιοπληροφορική. Γενικά πάντως, οι περισσότεροι εξακολουθούν να δέχονται ότι κυρίως το είδος των δεδομένων (μοριακά), το πλήθος τους (μεγάλο), αλλά και η μεθοδολογία ανάλυσης, ορίζουν το χώρο της Βιοπληροφορικής. Σε μια προσπάθεια να αποδώσουμε γραφικά κάτι τέτοιο, μπορούμε (αν και είναι σίγουρο ότι πολλοί δεν θα συμφωνήσουν) να θεωρήσουμε απλουστευτικά δύο άξονες: τον άξονα που περιέχει το είδος των δεδομένων υπό ανάλυση και τον άξονα που περιέχει το είδος της ανάλυσης και να παραστήσουμε εκεί το χώρο, που περιέχει τις δραστηριότητες της Βιοπληροφορικής αλλά και των συναφών επιστημών (Εικόνα 1.7).



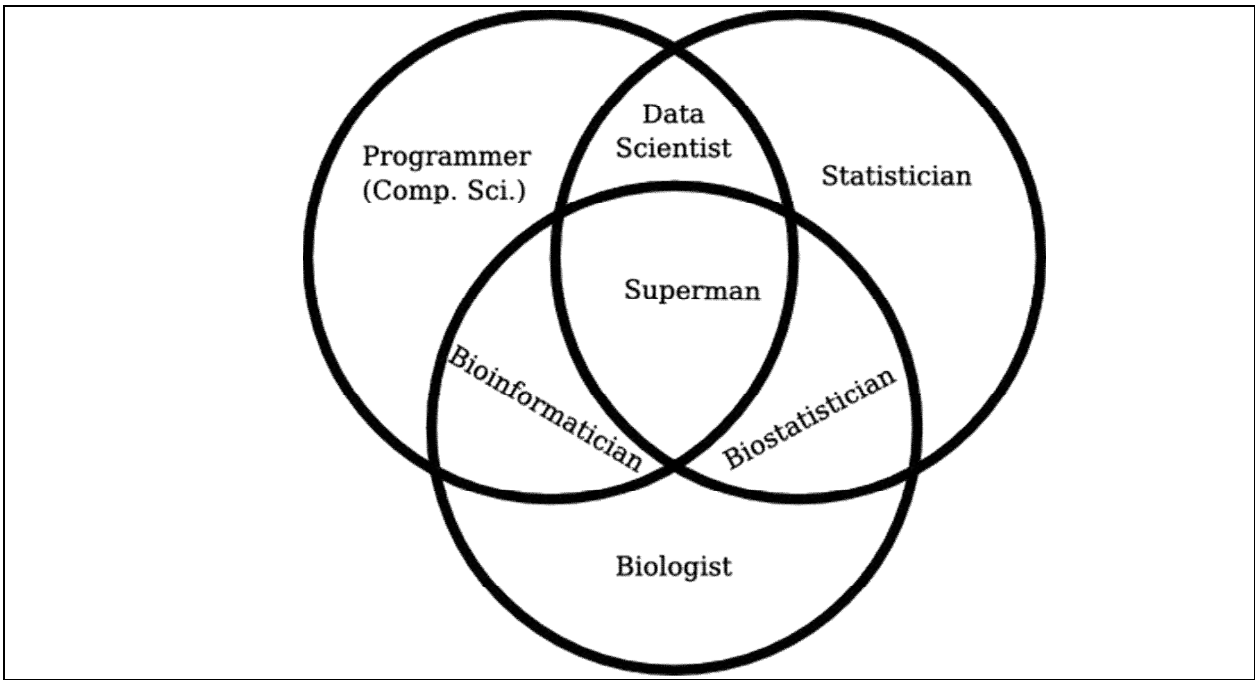
Εικόνα 1.7: *Μια προσπάθεια απεικόνισης της θέσης της βιοπληροφορικής σε σχέση με τις συγγενικές επιστήμες.*

Με μια τέτοια προσέγγιση, βλέπουμε τις περιοχές επαφής και αλληλεπικάλυψης, μικρές ή μεγάλες, με τις γειτονικές συναφείς ειδικότητες. Βλέπουμε λοιπόν ότι στην περιοχή των μικρών μορίων (φαρμάκων κλπ), η Βιοπληροφορική/Υπολογιστική Βιολογία εφάπτεται με την Υπολογιστική Χημεία, ενώ στην περιοχή της μοντελοποίησης, πολλές φορές ταυτίζεται με την Μαθηματική Βιολογία. Μια μεγάλη περιοχή επικάλυψης υπάρχει με την Ιατρική Πληροφορική στην περιοχή της μελέτης των κυτταρικών διεργασιών και των δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA, επικάλυψη που θα δούμε ότι επιβεβαιώνεται και από εμπειρικά δεδομένα της βιβλιογραφίας. Επίσης, η ίδια περιοχή, όπως και η περιοχή της ανάλυσης των γενετικών διαφορών των ατόμων και της συσχέτισης των πολυμορφισμών με ασθένειες (GWAS), αποτελεί περιοχή επικάλυψης αλλά και σύγκλισης της Βιοπληροφορικής με τη Βιοστατιστική (Molenberghs, 2005). Προφανώς, αν προσθέταμε και άλλους άξονες στο διάγραμμα αυτό, όπως π.χ. το σκοπό της μελέτης, θα μπορούσαμε να «διαχωρίσουμε» καλύτερα τις ειδικότητες. Όπως και να έχει, μιλάμε για μια υπεραπλουστευμένη ανάλυση, η οποία εν τούτοις μας δίνει κάποια χρήσιμα στοιχεία.

Η διεπιστημονικότητα της Βιοπληροφορικής, είναι επίσης μια έννοια με μεγάλες συζητήσεις και διχογνωμίες γύρω από αυτήν, καθώς επηρεασμένοι από την ανάγκη να δώσουν αποτελέσματα τα μεγάλα συνεργατικά προγράμματα (όπως το πρόγραμμα προσδιορισμού του ανθρώπινου γονιδιώματος), πολλοί, μεταξύ των οποίων και μεγάλοι ερευνητικοί οργανισμοί, δίνουν έμφαση στο σχηματισμό διεπιστημονικών ομάδων αντί στην εκπαίδευση διεπιστημονικών ατόμων. Όπως αναφέρει και ο Eddy (Eddy, 2005), πολλοί από εμάς που ασχολούμαστε με τη βιοπληροφορική για χρόνια, αντιμετωπίζουμε το ίδιο πρόβλημα που αντιμετώπιζαν και οι πρώτοι μοριακοί βιολόγοι: δεν μπορούμε να ταξινομηθούμε εύκολα στις «παραδοσιακές» ειδικότητες. Εγώ ας πούμε, σε μια παρόμοια διαδρομή με τον Eddy (αν και όχι τόσο επιτυχημένη, πρέπει να παραδεχτώ), σπούδασα Βιολογία και έκανα μεταπτυχιακό στη Βιοστατιστική (σε ένα διεπιστημονικό πρόγραμμα που αποτελούσε συνεργασία του Τμήματος Μαθηματικών και της Ιατρικής Σχολής), ενώ το διδακτορικό μου εκπονήθηκε σε Τμήμα Βιολογίας, αλλά με θέμα που ειναι ξεκάθαρα θέμα Βιοπληροφορικής («Πρόγνωση δομής και λειτουργίας μεμβρανικών πρωτεϊνών»). Το σύνολο του ερευνητικού μου έργου αφορά σε θέματα πρόγνωσης δομής και λειτουργίας πρωτεϊνών, κατασκευής βιολογικών βάσεων δεδομένων, ανάπυξη αλγορίθμων για HMM, ανάπτυξη στατιστικής μεθοδολογίας για μετα-ανάλυση γενετικών δεδομένων και εφαρμογές σε σημαντικές ασθένειες. Για να τα κάνω αυτά, χρησιμοποιώ τις βιολογικές γνώσεις μου, τις μαθηματικές μου γνώσεις στο σχεδιασμό αλγορίθμων και στατιστικών μεθόδων, ενώ στο τέλος κάποια από αυτά τα δημιουργήματά μου τα υλοποιώ σε κάποια γλώσσα προγραμματισμού. Παρόλο που σίγουρα υπάρχουν βιολόγοι με πολύ περισσότερες γνώσεις και δεξιότητες από μένα, στατιστικοί με καλύτερη θεωρητική κατάρτιση και προγραμματιστές πολύ πιο αποδοτικοί (αυτό είναι το πιο σίγουρο από όλα), είμαι σύμφωνα με τα περισσότερα κριτήρια ένας σχετικά επιτυχημένος βιοπληροφορικός. Παρόλα αυτά, σίγουρα θα βρεθούν βιολόγοι που θα αμφισβητήσουν ότι αυτό που κάνω είναι βιολογία («πάντα θα χρειάζεσαι ένα πείραμα», βλ. παρακάτω) ή όπως είπε και ο Eddy «I'm sure my union card has expired» (Eddy, 2005). Οι μαθηματικοί από την άλλη θα πουν ότι δεν έχω βασικό πτυχίο στα μαθηματικά ή τη στατιστική και ότι δεν έχω αποδείξει πολλά θεωρήματα, ενώ όσον αφορά την πληροφορική, τα πράγματα είναι χειρότερα: παρόλο που έχω διδάξει προγραμματισμό για χρόνια, έχω δημοσιεύσει αλγόριθμους και το λογισμικό μου χρησιμοποιείται από επιστήμονες σε όλον τον κόσμο, δεν έχω ούτε ένα σχετικό πτυχίο, ούτε καν ECDL (αυτό βέβαια, δεν ξέρω αν λέει περισσότερο κάτι για τη δική μου αξία ή για αυτήν του ECDL).

Το τελικό συμπέρασμα είναι ότι ναι μεν χρειάζεται σίγουρα διεπιστημονική συνεργασία διαφορετικών ειδικοτήτων, ιδιαίτερα στα πολύ μεγάλα και δύσκολα προβλήματα, αλλά σε έναν κλάδο που έχει αναπτύξει ήδη την δική του κουλτούρα, δυναμική και βιβλιογραφία, χρειάζεται πρώτα από όλα η εκπαίδευση διεπιστημονικών ατόμων, ατόμων που θα μπορούν να καταλάβουν τα βασικά από όλες τις «συνιστώσες» της βιοπληροφορικής, αλλά δεν είναι ανάγκη να είναι άριστοι και στις τρεις. Υπάρχουν πολλά ανέκδοτα περιστατικά, στα οποία μια διεπιστημονική ομάδα δεν μπόρεσε καν να συνενοηθεί στα βασικά (είναι σα να στέλνεις αντιπροσώπους στον ΟΗΕ, διπλωμάτες που δεν μιλάνε ξένη γλώσσα, όπως είπε πάλι πολύ εύστοχα ο Eddy). Μια ιστορία που μου έχουν διηγηθεί αφορούσε μια ομάδα στατιστικών που πήγε να συνεργαστεί σε ένα μεγάλο πρόγραμμα με μοριακούς βιολόγους. Στην πρώτη συνάντηση, οι βιολόγοι μίλαγαν επί μία και πλέον ώρα «*τα κατάλοιπα (σ.σ. residues) αυτό, τα κατάλοιπα το άλλο*» κ.ο.κ. Μετά από πολλή ώρα, κάποιος από τους στατιστικούς ρώτησε «*Ωραία όλα αυτά, αλλά δεν καταλαβαίνω τι εννοείτε. Κάνατε κάποια παλινδρόμηση; Από που προήλθαν αυτά τα κατάλοιπα;*» (σ.σ. residues ή residuals λέγονται τα κατάλοιπα της παλινδρόμησης, οι διαφορές δηλαδή που προκύπτουν αν από τις τιμές που προκύπτουν από το

μοντέλο της παλινδρόμησης, αφαιρεθούν οι παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής – οι βιολόγοι από την άλλη εννοούσαν απλά τα αμινοξικά κατάλοιπα της πρωτεϊνης). Καταλαβαίνουμε έτσι, ότι κάποιος που ξεκίνησε βιολόγος, δεν είναι απαραίτητο να είναι και ο καλύτερος προγραμματιστής, ούτε να αποδεικνύει θεωρήματα (αλλά είναι απαραίτητο να μπορεί να καταλάβει τι είναι ένας αλγόριθμος, και να μπορεί να γράψει 10 γραμμές κώδικα για να κάνει μια απλή ανάλυση). Όμοια, κάποιος που ξεκίνησε από την πληροφορική ή τα μαθηματικά, δεν είναι απαραίτητο να είναι γνώστης όλων των τελευταίων εξελίξεων και τεχνικών στη μοριακή βιολογία (αλλά πρέπει να μπορεί να καταλάβει τι είναι αμινοξύ, τι είναι κωδικόνιο, τι γονίδιο και τι πρωτεϊνη). Όπως αποτύπωσε με ένα διάγραμμα Venn ο Anthony Fejes, δεν είναι αναγκαίο να είναι κανείς υπεράνθρωπος για να είναι καλός βιοπληροφορικός (Εικόνα 1.8).

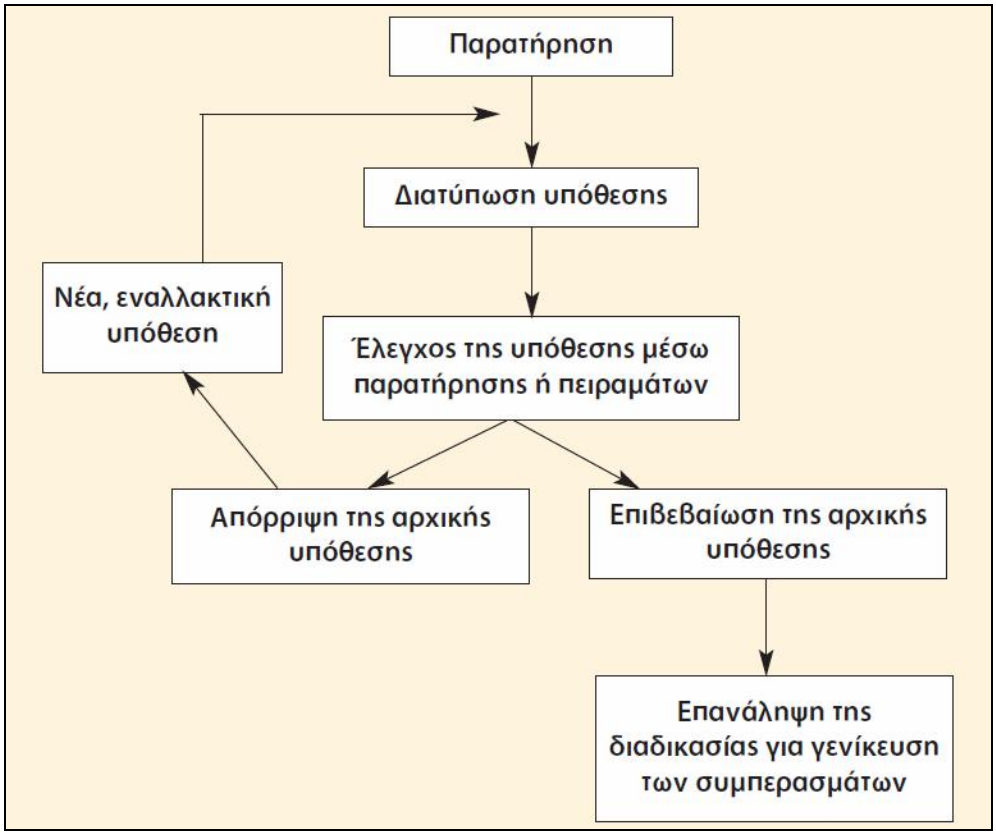


Εικόνα 1.8: *Από το* ***http://blog.fejes.ca/?p=2418***

Ένα σημείο που πρέπει να διευκρινιστεί παράλληλα με τη διεπιστημονικότητα, είναι και το ίδιο το επιστημολογικό καθεστώς της Βιοπληροφορικής (C. Ouzounis, 2002). Το θέμα μπορεί να γίνει κατανοητό, με ένα απλό παράδειγμα, που όμως δεν απέχει και πολύ από την πραγματικότητα. Ας υποθέσουμε λοιπόν ότι έχουμε έναν άψογα εκπαιδευμένο μοριακό βιολόγο και τον στείλουμε να σπουδάσει είτε σε προπτυχιακό είτε σε μεταπτυχιακό επίπεδο, πληροφορική. Αυτό θα τον κάνει αυτομάτως βιοπληροφορικό; Η απάντηση είναι ένα κατηγορηματικό όχι. Φυσικά, το ίδιο ισχύει, και ίσως για την ακρίβεια να είναι και χειρότερη η κατάσταση, για κάποιον μαθηματικό ή πληροφορικό που θα κληθεί να εκπαιδευτεί στη βιολογία. Ακριβώς όπως στην περίπτωση της διεπιστημονικής ομάδας ατόμων που είδαμε παραπάνω, έτσι και εδώ, υπάρχουν πολλά περισσότερα από μια απλή εκπαίδευση στις βασικές ειδικότητες, όσο καλή και επιτυχημένη και αν είναι αυτή. Η βιοπληροφορική, πέρα από το συνδυασμό γνώσεων από τις βασικές «συνιστώσες» της, έχει και το δικό της επιστημονικό βάθος. Έχει τη δική της ορολογία, τα δικά της προβλήματα, τις δικές της μεθοδολογίες, την ξεχωριστή της κουλτούρα, αλλά και τη δική της βιβλιογραφία που όπως είδαμε πηγαίνει 50 χρόνια πίσω. Όλοι οι παραπάνω παράγοντες, έχουν φυσικά τις ρίζες τους στις βασικές «συνιστώσες» (βιολογία, μαθηματικά, πληροφορική), οι οποίες μπορούν να εντοπιστούν ιστορικά, αλλά το γεγονός παραμένει, ότι η βιοπληροφορική διεκδικεί πλέον, και κατά τη γνώμη μου το έχει πετύχει, καθεστώς αυτόνομης επιστημονικής οντότητας, αναγνωρίζοντας φυσικά τις συγγένειες και τις εξαρτήσεις με τα άλλα πεδία. Επιπλέον, γίνεται κατανοητό με το παραπάνω παράδειγμα, ότι η πιο σωστή ονομασία θα ήταν Υπολογιστική Βιολογία, για να τονιστεί ακριβώς η έμφαση στο αντικείμενο της μελέτης (τα βιολογικά συστήματα), κατά ανάλογο τρόπο με τη Μοριακή Βιολογία. Αντίθετα, το όνομα Βιοπληροφορική, παραπέμπει σε άλλους κλάδους όπως π.χ. τη Βιοστατιστική, η οποία όμως είναι κατά βάση ειδικότητα της

στατιστικής (με τις όποιες ιδιαιτερότητές της) ή τη Γεωπληροφορική η οποία είναι απλά η εφαρμογή μεθόδων πληροφορικής σε προβλήματα χωροταξίας και χαρτογράφησης.

Το παραπάνω, είναι ένα κομβικό σημείο στην κατανόηση του επιστημολογικού πλαισίου της Βιοπληροφορικής και την αντιμετώπισή της ως Υπολογιστική Βιολογία, και είναι κάτι που πολλές φορές δεν γίνεται κατανοητό ούτε από τους παραδοσιακούς βιολόγους, οι οποίοι επίσης αντιμετωπίζουν τη βιοπληροφορική απλά σαν μια «εφαρμογή μεθόδων». Ένα κλασικό παράδειγμα, βρίσκεται σε μια φράση που όλοι όσοι ασχολούμαστε με τη βιοπληροφορική έχουμε λίγο πολύ ακούσει («*εντάξει, καλά είναι αυτά που μας λες, οι προγνώσεις, οι στοιχίσεις και όλα αυτά, αλλά πάντα θα χρειάζεσαι το πείραμα*»). Αυτή η φράση, μπορεί φυσικά να περιέχει αλήθεια σε πολλές περιπτώσεις, δεν μπορεί όμως να έχει καθολική εφαρμογή και δείχνει απλά μια προσκόλληση σε μια υπερβολικά απλουστευμένη, απλοϊκή και τελικά στρεβλή εκδοχή αυτού που ακόμα και στα σχολικά βιβλία βιολογίας αναφέρεται ως «επιστημονική μέθοδος». Το μοντέλο αυτό, καθώς είναι επηρεασμένο από το θετικισμό αλλά και τη διαψευσιοκρατία, θεωρείται από τις σύγχρονες προσεγγίσεις περί φιλοσοφίας της επιστήμης ως μη επαρκές θεωρητικά, αλλά παρόλα αυτά έχει δώσει μεγάλες επιτυχίες στη σύγχρονη βιολογία και καλώς χρησιμοποιείται. Τα θεωρητικά επιστημολογικά προβλήματα αυτής της προσέγγισης αφορούν κυρίως την εξάρτηση της παρατήρησης από τη θεωρία, την επισφάλεια όλων των πειραμάτων, αλλά και την τελική αδυναμία να δοθεί μια ξεκάθαρη απάντηση και μια μέθοδος για το πώς παράγεται τελικά μια ολοκληρωμένη θεωρία (Chalmers, 1999).



Εικόνα 1.9: *Σχήμα που απεικονίζει την επιστημονική μέθοδο (Μαυρικάκη, Γκούβρα, & Καμπούρη, 2014)*

Για αρχή, η παρατήρηση δεν είναι κάτι ουδέτερο και εξαρτάται από τη θεωρία και το ολοκληρωμένο σύστημα αξιών μέσα στο οποίο πραγματοποιείται.. Λίγες φορές στις σύγχρονες επιστήμες η παρατήρηση είναι οπτική (αλλά ακόμα και τότε είναι επισφαλής), ενώ στις περισσότερες των περιπτώσεων η κατανόηση του τι παρατηρήθηκε απαιτεί την αποδοχή ενός συνόλου κανόνων, τεχνικών, πορισμάτων κ.ο.κ. Η «παρατήρηση» ότι μια συγκεκριμένη πρωτεΐνη έχει μια συγκεκριμένη τρισδιάστατη μορφή, απαιτεί την αποδοχή της τεχνικής της κρυσταλλογραφίας, της περίθλασης ακτίνων Χ, της επίλυσης του προβλήματος φάσης, την ανασύσταση της δομής κ.ο.κ. Το ίδιο φυσικά ισχύει και για άλλες «παρατηρήσεις», οι οποίες στην

ουσία είναι πειράματα οι ίδιες (η αλληλούχιση, η PCR, η ηλεκτροφόρηση κ.ο.κ.). Το ένα πρόβλημα με αυτή τη θεώρηση, είναι ότι δεν μπορείς να παρατηρήσεις εύκολα κάτι που δεν ταιριάζει στο δικό σου σύστημα. Υπάρχουν αρκετά τέτοια παραδείγματα στην ιστορία της μοριακής βιολογίας (π.χ. τα δεδομένα κρυσταλλογραφίας ακτίνων Χ με τα οποία οι Watson και Crick προσδιόρισαν τη δομή του DNA ήταν διαθέσιμα από καιρό αλλά δεν μπορούσαν να αξιοποιηθούν). Το άλλο πρόβλημα, είναι ότι τα δεδομένα αυτά είναι από τη φύση τους επισφαλή. Οι τεχνικές έχουν σφάλματα, υπόκεινται σε πειραματικό λάθος και είναι κάθε άλλο παρά τέλειες. Υπάρχουν επίσης πολλά παραδείγματα όπου ένα πείραμα δεν εκτελέστηκε σωστά (π.χ. έγινε μια επιμόλυνση της καλλιέργειας) ή, ακόμα χειρότερα, ένα μικρό σφάλμα στην όλη διαδικασία (π.χ. ένα μικρό τυπογραφικό λάθος σε ένα από τα πολλά προγράμματα κρυσταλλογραφίας) οδήγησε σε τρισδιάστατες δομές που ήταν τελείως λάθος.

Το βασικό όμως πρόβλημα, για να επιστρέψουμε στο αρχικό μας ερώτημα, δεν είναι όλα τα παραπάνω (καθώς η βιοπληροφορική εντάσσεται ξεκάθαρα στον κορμό των βιολογικών και των άλλων θετικών επιστημών), αλλά η στρεβλή και απλοϊκή αντιμετώπιση του τελευταίου βήματος, του τρόπου δηλαδή παραγωγής της θεωρίας και εξαγωγής των γενικών νόμων. Για παράδειγμα, κάποια φαινόμενα είναι απλά, με την έννοια ότι επιδέχονται μια απλή και ξεκάθαρη απάντηση. Έτσι, όταν ξεκίνησε η διερεύνυση του γενετικού κώδικα, υπήρξε η παρατήρηση ότι τη τριπλέτα UUU κωδικοποιεί το αμινοξύ Φενυλαλανίνη (Phe). Αυτή η παρατήρηση, δεν χώραγε αμφισβήτηση, ενώ με τα αντίστοιχα (και ιδιαίτερα έξυπνα μπορούμε να πούμε) πειράματα για τις υπόλοιπες τριπλέτες αποκωδικοποιήθηκε ολόκληρος ο γενετικός κώδικας με τρόπο αδιαμφισβήτητο. Ο γενετικός κώδικας σύμφωνα με όσα είναι γνωστά από τη δεκαετία του 1960, είναι μια απλή συνάρτηση μίας μεταβλητής, μια απεικόνιση του συνόλου των κωδικονίων στο σύνολο των αμινοξέων, με την οποία κάθε μέλος του πρώτου συνόλου αντιστοιχείται σε ένα μόνο μέλος του δεύτερου συνόλου (μια συνάρτηση όμως που δεν είναι «ένα προς ένα», και κατά συνέπεια δεν είναι αντιστρέψιμη). Τι γίνεται όμως με περιπτώσεις στις οποίες τα πράγματα δεν είναι τόσο ξεκάθαρα; Σε περιπτώσεις που οι αντιστοιχίσεις δεν είναι τόσο απλές; Στην περίπτωση λ.χ. της δομής των πρωτεϊνών, παρόλα τα πειράματα και τα θεωρητικά επιχειρήματα που δείχνουν ξεκάθαρα ότι η αλληλουχία καθορίζει τη δομή, δεν υπάρχει κάποιος ξεκάθαρος κώδικας, κάποιος κανόνας που να λέει ότι η αλληλουχία των X-Y-Z αμινοξέων θα έχει πάντα τη δομής α-έλικας, ενώ η αλληλουχία των Α-Β-Γ θα έχει τη δομή β-πτυχωτής επιφάνειας. Εδώ, το πρόβλημα προκύπτει εγγενώς ασαφές και πολυδιάστατο και όσα δεδομένα και αν συλλέξουμε, όσα πειράματα προσδιορισμού δομών και αν κάνουμε, δεν θα μπορέσουμε ποτέ να καταλήξουμε (τουλάχιστον με τον απλοϊκό τρόπο που είδαμε παραπάνω) σε τόσο απλά διατυπωμένους καθολικούς νόμους.

Σε αυτό το σημείο έρχεται να συμβάλει η Βιοπληροφορική, η οποία χρησιμοποιώντας τα υπάρχοντα πειραματικά δεδομένα (τα οποία μπορεί να αντιπροσωπεύουν χιλιάδες ανθρωποώρες επίπονης δουλειάς των εργαστηριακών βιολόγων), χρησιμοποιεί τεχνικές της στατιστικής, των μαθηματικών και της πληροφορικής, με σκοπό να εξάγει ένα γενικό νόμο ή έστω κάποιους κανόνες που να τον προσεγγίζουν. Εισάγει δηλαδή τη μαθηματικοποίηση και την ποσοτικοποίηση των βιολογικών φαινομένων, μια προσέγγιση που όπως είδαμε δεν είναι καθόλου νέα στη βιολογία. Στο παράδειγμα της δομής των πρωτεϊνών, η υπολογιστική μελέτη των χιλιάδων τρισδιάστατων δομών των πρωτεϊνών, έχει καταλήξει σε κάποιους γενικούς νόμους, οι οποίοι δεν αποτελούν -και δεν θα μπορούσαν να αποτελέσουν ποτέ- το αποτέλεσμα κάποιου συγκεκριμένου «πειράματος». Τέτοιοι νόμοι, είναι οι: α) οι πρωτεϊνικές δομές συντηρούνται περισσότερο από τις πρωτεϊνικές αλληλουχίες και β) οι περισσότερες σημειακές μεταλλάξεις στις πρωτεΐνες συμβαίνουν στην επιφάνειά τους παρά στο εσωτερικό της δομής. Αυτοί οι νόμοι μπορεί να θεωρηθούν ως γενικότεροι νόμοι των βιολογικών επιστημών, καθώς έχουν γενικότερες συνέπειες σε πολλά πεδία και δίνουν άμεσες απαντήσεις σε πολλά πρακτικά ερωτήματα. Για παράδειγμα, αν εντοπίσουμε μια πρωτεΐνη με 99% ομοιότητα με μια πρωτεΐνη γνωστής δομής και λειτουργίας, οι νόμοι αυτοί μας λένε ξεκάθαρα και με μεγάλη βεβαιότητα ότι τα αμινοξικά κατάλοιπα που διαφέρουν θα βρίσκονται στην επιφάνεια του μορίου, θα έχουν ελάχιστη επίδραση στην τρισδιάστατη δομή και κατά κανόνα δεν θα επηρεάζουν σημαντικά τη γενική βιολογική λειτουργία. Επιπλέον δε, οι αλγόριθμοι πρόγνωσης της δομής, ακόμα και αν δεν δίνουν κάποιον ξεκάθαρο κανόνα, μπορούν να κάνουν προβλέψεις για τη δομή μιας πρωτεΐνης, προβλέψεις για τις οποίες ξέρουμε με μεγάλη αξιοπιστία τι ποσοστό επιτυχίας αναμένουμε να έχουν. Τέτοιοι νόμοι και η διαδικασία με την οποία προέρχονται (η Βιοπληροφορική δηλαδή) μπορούν από τη μία μεριά να επιταχύνουν τη βιολογική έρευνα οργανώνοντας τον όγκο των πειραματικών δεδομένων, αλλά και αντικαθιστώντας από την άλλη, όπου αυτό είναι δυνατό, τα επιπλέον πειράματα αποκλείοντας μη πιθανές εκδοχές. Εάν λάβουμε υπ' όψιν όλα τα παραπάνω, δεν συνεπάγεται ότι πρέπει να μειώσουμε την αξία του πειραματισμού και της πειραματικής βιολογίας (αν μη τι άλλο, αν δεν υπήρχαν τα τεράστια αποθέματα πειραματικών δεδομένων, δεν θα υπήρχε

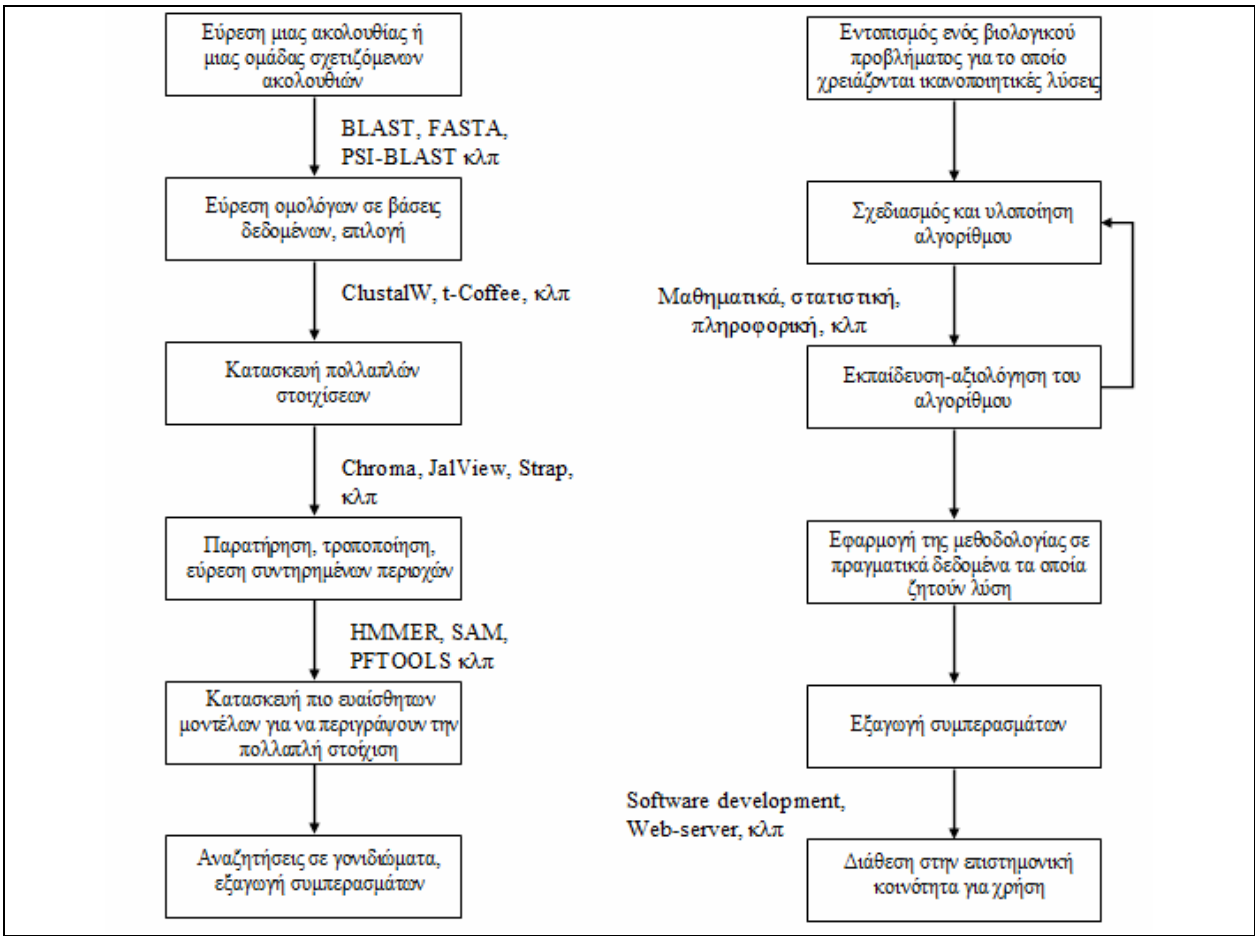
και βιοπληροφορική). Αυτό που πρέπει να γίνει, είναι να αναγνωριστεί η Βιοπληροφορική και η Υπολογιστική Βιολογία σαν ένας αυτόνομος κλάδος των βιολογικών επιστημών με τον ίδιο ή περίπου τον ίδιο τρόπο με τον οποίο οι φυσικοί έχουν αποδεχθεί την υπολογιστική φυσική και οι χημικοί την υπολογιστική χημεία.

Η παραπάνω λανθασμένη αντίληψη περί βιοπληροφορικής, αποτελεί έναν από τους «μύθους περί βιοπληροφορικής», που ανέλυσε ο Χρήστος Ουζούνης (C. Ouzounis, 2000). Ένας άλλος μύθος που σχετίζεται βέβαια με αυτή τη λάθος θεώρηση, είναι ότι «η βιοπληροφορική είναι μια νέα τεχνολογία». Αφενός μεν, όπως είδαμε προηγουμένως, αν και ο όρος είναι νέος, το αντικείμενο της υπολογιστικής βιολογίας και βιοπληροφορικής έχει βάθος δεκαετιών. Αφετέρου δε, με όσα αναφέραμε ήδη, θα πρέπει να γίνεται κατανοητό ότι δεν είναι «τεχνολογία». Τεχνολογία, τουλάχιστον στο βαθμό που μας αφορά σχετικά με τη συνάφειά της με τη βιολογία, είναι για παράδειγμα, οι μικροσυστοιχίες, το pyrosequencing, οι σχεσιακές βάσεις δεδομένων ή μια νέα γλώσσα προγραμματισμού. Αντίθετα, η βιοπληροφορική είναι ένα εκτεταμένο σύνολο εννοιών και μεθοδολογιών, που όπως είδαμε παραπάνω συνιστά ξεχωριστή επιστημονική ειδικότητα. Αυτή η παρανόηση προέρχεται από την αποσπασματική εικόνα που έχουν πολλοί, ειδικά στο χώρο της βιολογίας. Για παράδειγμα, ένας ερευνητής που ασχολείται για χρόνια με τη γονιδιακή έκφραση, έρχεται για πρώτη φορά σε επαφή με την τεχνολογία των μικροσυστοιχιών η οποία απογειώνει τη δουλειά του. Μαζί όμως με τα πανάκριβα μηχανήματα, τα αντιδραστήρια, και τον υπόλοιπο εξοπλισμό, βλέπει και έναν 'περίεργο τύπο' να «πατάει τα κουμπιά» και να βγάζει αποτέλεσμα. Είναι λογικό κατά κάποιον τρόπο να υποθέσει ότι αυτό μόνο είναι η βιοπληροφορική, μια τεχνολογική πλατφόρμα για να διευκολύνει τη δουλειά του, όπως οι Η/Υ, τα λειτουργικά συστήματα και ο επεξεργαστής κειμένου στον οποίο θα γράψει την εργασία του. Αγνοεί όμως, αφενός μεν το θεωρητικό υπόβαθρο που υπάρχει πίσω από όλες τις πλατφόρμες λογισμικού που χρησιμοποιεί ο «τεχνικός» και την ειδική γνώση που χρειάζεται για την κατανόηση των αποτελεσμάτων, αφετέρου δε το γεγονός ότι υπάρχει και άλλου είδους βιοπληροφορική. Υπάρχουν αυτοί που κατασκευάζουν τους αλγόριθμους, αυτοί που αποδεικνύουν τα θεωρήματα, αυτοι που σχεδιάζουν το λογισμικό, κ.ο.κ. και όλα αυτά, για ένα μεγάλο εύρος ερευνητικών ερωτημάτων, όχι μόνο για τις μικροσυστοιχίες που είναι (ή ήταν) της μόδας (π.χ. στοίχιση αλληλουχιών, φυλογενετική ανάλυση, πρόγνωση δομής, μοντελοποίηση με βάση την ομολογία, κατασκευή βιολογικών βάσεων δεδομένων κ.ο.κ.). Μια τέτοια αντίληψη υπάρχει δυστυχώς παγκοσμίως σε μερίδα των βιολόγων, και όπως έγραφε με παράπονο ο Edgar Wingender: «*Thus, scientific articles publishing experimental findings which have been evaluated using computational tools, very often give credit to them in the Methods or Results sections with phrases such as "Computer analysis revealed that ...", without any appropriate reference. In contrast, any experimental methodology used is extensively explained in these papers, down to the detailed listing of buffer systems, voltage/current conditions of the electrophoresis systems etc*» (Wingender, 1998).

Η λάθος αυτή εντύπωση που δημιουργείται σε πολλούς, οδηγεί και σε έναν άλλο πολύ διαδεδομένο μύθο, αυτόν που λέει ότι «η βιοπληροφορική είναι εύκολη», με επακόλουθο το ότι «ο καθένας μπορεί να το κάνει» και «οι εργασίες οι δικές σας βγαίνουν εύκολα, πατάτε δυο κουμπιά και βγάζετε paper». Παλι, μια τέτοια θεώρηση είναι λανθασμένη, αν και πρέπει να αναγνωρίσουμε στους πειραματικούς βιολόγους ότι ειδικά στην Ελλάδα η υψηλού επιπέδου έρευνα στις μοριακές επιστήμες είναι ακριβή, αλλά το κυριότερο, δύσκολη καθώς, εκτός από την εξασφάλιση των κονδυλίων, πρέπει κανείς να αναμετρηθεί και με τη γραφειοκρατία, να ασχοληθεί με διαγωνισμούς και προμήθειες και ακόμα και αν είναι τυχερός με όλα αυτά, μπορεί να περιμένει μήνες για να παραλάβει τα ακριβά του αντιδραστήρια και τα μηχανήματα. Παρόλα αυτά, η κατάσταση για κάποιον που έχει μια εποπτική εικόνα της βιοπληροφορικής, δεν είναι στο σύνολο της, δραματικά καλύτερη. Ναι, υπάρχουν κάποια ερευνητικά πεδία που απαιτούν λιγότερη επένδυση σε υλικό και λογισμικό, αλλά στις περισσότερες περιπτώσεις απαιτούνται ισχυροί υπολογιστές και μεγάλοι αποθηκευτικοί χώροι. Όπως και να έχει όμως, όλες οι εργασίες βιοπληροφορικής απαιτούν εξειδικευμένο προσωπικό υψηλών προσόντων, κάτι το οποίο δεν είναι ούτε εύκολο να βρεθεί, αλλά ούτε και «χαμηλού κόστους». Κατά συνέπεια, το «ο καθένας μπορεί να το κάνει» δεν ευσταθεί, γιατί αν μπορούσε, θα το είχε κάνει. Δεν μπορεί ο καθένας να φτιάξει έναν επιτυχημένο αλγόριθμο πρόγνωσης, γιατί μια τέτοια διαδικασία απαιτεί ακριβώς τις εξειδικευμένες γνώσεις που απαιτεί η διεπιστημονική προσέγγιση που αναφέραμε, ούτε μπορεί ο καθένας να κάνει μια συγκριτική ανάλυση όλων των γνωστών γονιδιωμάτων, γιατί κάτι τέτοιο απαιτεί επιπλέον και μεγάλη υπολογιστική ισχύ και αποθηκευτικό χώρο. Αλλά ούτε και όταν κάποιος έχει τη δυνατότητα να φτιάξει έναν τέτοιον αλγόριθμο ή να κάνει μια τέτοια ανάλυση, αυτό σημαίνει ότι αυτά γίνονται «γρήγορα». Τέτοιες δραστηριότητες, απαιτούν προσεκτικό σχεδιασμό και πειραματισμό, δοκιμή και σφάλμα, διαδικασίες

δηλαδή επίπονες και χρονοβόρες. Είναι δηλαδή, από όλες τις απόψεις, πραγματικά πειράματα και σαν τέτοια θα έπρεπε να αντιμετωπίζονται.

Σε αυτό το σημείο, πρέπει να κάνουμε όμως και έναν επιπλέον διαχωρισμό των δραστηριοτήτων βιοπληροφορικής. Οι δραστηριότητες που αναφέραμε προηγουμένως, η κατασκευή μεθόδου πρόγνωσης, η μεγάλης κλίμακας υπολογιστικές αναλύσεις, ο σχεδιασμός αλγορίθμων και λογισμικού κ.ο.κ. ανήκουν στην κατηγορία δραστηριοτήτων που όντως απαιτούν μια μεγάλη εξειδίκευση και δεν μπορούν να γίνουν από τον καθένα. Σήμερα όμως, με την πρόοδο που έχει επιτευχθεί σε όλους τους τομείς, υπάρχει και μια μεγάλη ομάδα δραστηριοτήτων που μπορεί να τις επιτελέσει ο καθένας, και μάλιστα θα έλεγα ότι είναι απαραίτητο να μπορεί να τις πραγματοποιεί ο καθένας που ασχολείται με τη βιολογική έρευνα. Η τεράστια ανάπτυξη της υπολογιστικής βιολογίας, όπως είδαμε, έχει φέρει τη χρήση αλγοριθμικών και υπολογιστικών εργαλείων στην καθημερινότητα του βιολόγου (δεν υπάρχει βιολόγος που να μην έχει χρειαστεί να χρησιμοποιήσει το BLAST). Έτσι θα λέγαμε ότι υπάρχουν διάφορες κατατάξεις στον τρόπο που ένας ερευνητής χρησιμοποιεί και εμπλέκεται με δραστηριότητες βιοπληροφορικής. Στην πρώτη κατηγορία, έχουμε τις απλές αναλύσεις που αφορούν τη χρήση λογισμικού (στοίχιση, πολλαπλή στοίχιση, μέθοδος πρόγνωσης, αναζήτηση σε βάσεις δεδομένων κ.ο.κ.). Αυτές τις δραστηριότητες θα μπορούσε και θα έπρεπε να μπορεί να της φέρει σε πέρας ο κάθε βιολόγος ανεξάρτητα της ειδικότητας του και του αντικειμένου της έρευνάς του, αλλά από την εμπειρία μας έχουμε δει ότι η έλλειψη θεωρητικής κατανόησης για το τι ακριβώς κάνει η κάθε μέθοδος, οδηγεί σε πολλά προβλήματα (λάθος χρήση μιας μεθόδου). Στην επόμενη κατηγορία, μπορούμε να κατατάξουμε αυτούς που χρησιμοποιούν κατά κύριο λόγο τέτοια έτοιμα εργαλεία και αλγόριθμους για να πραγματοποιήσουν σύνθετες αναλύσεις και να απαντήσουν σε κάποιο βιολογικό ερώτημα. Τα τελευταία χρόνια, η εμφάνιση των δεδομένων γονιδιακής έκφρασης, τα δεδομένα αλληλούχισης νέας γενιάς κ.ο.κ. έχουν αυξήσει αυτού του είδους τις αναλύσεις και την ανάγκη για άτομα που να μπορούν να τις φέρουν σε πέρας. Το βασικό όμως χαρακτηριστικό αυτής της ομάδας είναι ότι δεν αναπτύσσει αλγόριθμους, ούτε λογισμικό. Τέλος, υπάρχουν και αυτοί που εστιάζοντας σε κάποιο συγκεκριμένο πρόβλημα αναπτύσσουν και αλγόριθμους και λογισμικό. Οι αλγόριθμοι μπορεί να είναι αλγόριθμοι πρόγνωσης, στοίχισης ή οποιαδήποτε άλλη κατηγορία από αυτές που έχουμε αναλύσει.



Εικόνα 1.10: *Αριστερά, η διεργασία που επιτελεί ένας χρήστης βιοπληροφορικής. Δεξιά, η διεργασία που επιτελεί ένας βιοπληροφορικός*

Υπάρχει στη βιβλιογραφία μια έντονη διχογνωμία για το πώς πρέπει να ονομαστεί ο επιστήμονας της κάθε κατηγορίας, ειδικά για τις κατηγορίες 2 και 3 (στην 1 είναι οι βιολόγοι όπως είπαμε). Έτσι, έχουν προταθεί οι όροι «bioinformatician» για τους επιστήμονες της κατηγορίας 2 και «bioinformaticist» για αυτούς της κατηγορίας 3, αλλά η διάκριση δεν έχει γίνει αποδεκτή και ο τελευταίος όρος δεν χρησιμοποιείται πολύ. Άλλοι χρησιμοποιούν τον όρο «bioinformatics scientist» και «bioinformatics engineer» για τις κατηγορίες 2 και 3 αντίστοιχα (Welch et al., 2014), αλλά και αυτή η προσέγγιση προσωπικά δεν μου φαίνεται σωστή, γιατί μπορεί να δημιουργήσει την εντύπωση (ειδικά στην Ελλάδα), ότι πρέπει υποχρεωτικά οι επιστήμονες να είναι μηχανικοί, δηλαδή απόφοιτοι πολυτεχνείου. Ίσως ένας περιφραστικός ορισμός να είναι αναγκαίος, πάντα ανάλογα με το περιεχόμενο και την περίσταση. Όπως και να έχει όμως, η άποψή μου είναι ότι μια σωστή διεπιστημονική εκπαίδευση, ακόμα και στο πλαίσιο του βασικού πτυχίου (αλλά σίγουρα και στο πλαίσιο ενός μεταπτυχιακού προγράμματος), θα επέτρεπε στους βιολόγους να μπορούν να αποδίδουν τα μέγιστα, ακόμα και στην περιοχή της εξειδίκευσης τους. Υπάρχουν ένα σωρό παραδείγματα υπολογιστικών αναλύσεων οι οποίες θα μπορούσαν να είχαν γίνει ακόμα και από έναν «παραδοσιακό» βιολόγο (με την έννοια ότι δεν χρειάζεται ειδικές γνώσεις προγραμματισμού), όπως η κατασκευή τρισδιάστατων μοντέλων πρωτεϊνών, ο χαρακτηρισμός μιας πρωτεϊνικής οικογένειας και η εύρεση μακρινών ομολόγων, αλλά αφέθηκαν στους «ειδικούς», τους βιοπληροφορικούς. Επιπλέον, μια μεταπτυχιακή εκπαίδευση στη χρήση των βασικών εργαλείων, ειδικά όσον αφορά τις νέες τεχνολογίες αλληλούχισης και γονιδιακής έκφρασης, θα μπορούσε να αποτελέσει και μια επαγγελματική διέξοδο με περισσότερες προοπτικές, καθώς οι τεχνολογίες αυτές χρησιμοποιούνται πλέον στα περισσότερα εργαστήρια μοριακής βιολογίας αλλά και σε νοσοκομεία και διαγνωστικά κέντρα. Φυσικά, σε αυτή την κατηγορία δεν είναι απαραίτητο να εντάσσονται μόνο βιολόγοι (αν και θα ήταν ίσως πιο εύκολο για αυτούς), αλλά και επιστήμονες άλλων ειδικοτήτων όπως πληροφορικοί και στατιστικοί, αφού θα έχουν περάσει πρώτα από κάποιου είδους εκπαίδευση. Οι επιστήμονες της 3ης

κατηγορίας, συνιστούν την πιο ετερογενή ομάδα, καθώς σε αυτό το πρότυπο μπορεί να ταιριάζουν επιστήμονες με διαφορετικό προφίλ, από μοριακούς βιολόγους και φυσικούς, μέχρι θεωρητικούς πληροφορικούς και μαθηματικούς που ασχολήθηκαν με ένα συγκεκριμένο πρόβλημα της υπολογιστικής βιολογίας και εστιάζουν στο πώς θα αναπτύξουν αλγόριθμους και λογισμικό για την αντιμετώπισή του.

Αφού είδαμε την ιστορική διαδρομή της βιοπληροφορικής και τις θεωρητικές αναλύσεις που δικαιολογούν τη διεπιστημονικότητα του κλάδου, αναγκαστικά καταλήγουμε στην κουβέντα για την εκπαίδευση των βιοπληροφορικών. Γενικά, υπάρχει μεγάλη συζήτηση στη βιβλιογραφία για το ποια θα πρέπει να είναι η κατάλληλη εκπαίδευση, για το πώς θα πρέπει να δομούνται τα διάφορα προγράμματα σπουδών ειδικά όταν θα απευθύνονται σε διαφορετικό ακροατήριο, για το πώς θα επιχυγχάνεται η διεπιστημονική προσέγγιση, αλλά και για το πώς θα ενσωματωθούν τα μαθήματα βιοπληροφορικής στα βασικά προγράμματα σπουδών των βιοεπιστημών και της ιατρικής (Altman, 1998; Ditty et al., 2010; Floriano, 2008; Honts, 2003; Searls, 2012; Welch, et al., 2014; Yan, Ban, & Tan, 2014). Αυτή η διεπιστημονική εκπαίδευση, η οποία είχε ξεκινήσει ήδη από τη δεκαετία του 1990 στο εξωτερικό, έχει αρχίσει να γίνεται αποδεκτή σταδιακά και στην Ελλάδα. Προγράμματα Μεταπτυχιακών Σπουδών (ΠΜΣ) έχουν ήδη ιδρυθεί και λειτουργούν εδώ και χρόνια (βλ. παρακάτω), τα οποία δέχονται αποφοίτους όλων των παραπάνω κατηγοριών (βιολόγους, γιατρούς, μαθηματικούς, στατιστικούς, πληροφορικούς), ενώ το πρόγραμμα σπουδών τους διαιρείται με άξονα τις βασικές αρχές που περιέγραψε πρώτος ο Altman: βασική βιολογία, βιοστατιστική, προγραμματισμός και βασικά στοιχεία επιστήμης υπολογιστών και τέλος ειδικές γνώσεις της βιοπληροφορικής (ανάλυση ακολουθιών, ανάλυση δομών, βιολογικές βάσεις δεδομένων, κ.ο.κ.). Στο προπτυχιακό επίπεδο, όπως θα δούμε παρακάτω, τα πράγματα είναι πιο σύνθετα, καθώς η βιοπληροφορική πρέπει να ενσωματωθεί σε ένα υπάρχον πρόγραμμα σπουδών. Έτσι, τα περισσότερα τμήματα βιολογίας έχουν ανταποκριθεί στις ανάγκες της εποχής εντάσσοντας στο πρόγραμμά τους κάποιο μάθημα βιοπληροφορικής, αλλά το ίδιο δε συμβαίνει για τα περισσότερα τμήματα πληροφορικής ή μηχανικών Η/Υ. Το Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Θεσσαλίας, είναι κατά κάποιον τρόπο μοναδικό στον τομέα αυτό, καθώς παρέχει και σε προπτυχιακό επίπεδο τη διεπιστημονικότητα που χρειάζεται, καθώς στον κορμό των μαθημάτων πληροφορικής που είναι κοινός με τα περισσότερα τμήματα πληροφορικής, παρέχει και μια σειρά μαθημάτων στη Βιολογία, Βιοχημεία, Γενετική, Φυσιολογία, και φυσικά Βιοπληροφορική, Βιοστατιστική και Ιατρική Πληροφορική.

1.3. Η κατάσταση στον κόσμο

Την τελευταία δεκαετία έχουν γίνει πολλές επιστημομετρικές μελέτες με σκοπό να μελετήσουν ειδικά την επιστήμη της βιοπληροφορικής και συγκεκριμένα την επιστημονική βιβλιογραφία του χώρου. Άλλες δίνουν έμφαση στις ερευνητικές κατευθύνσεις που ακολουθεί ο κλάδος παγκοσμίως, άλλες επιχειρούν να σκιαγραφήσουν το τοπίο εστιάζοντας στα επιδραστικά περιοδικά, στους συγγραφείς και στα ερευνητικά ιδρύματα που εμπλέκονται στο χώρο, ενώ άλλες επιχειρούν να εστιάσουν στις ομοιότητες και τις διαφορές με άλλους συγγενείς κλάδους, κυρίως με την Ιατρική Πληροφορική. Σε όλες τις περιπτώσεις ένα βασικό πρόβλημα που αντιμετωπίζει μια τέτοια μελέτη είναι στο πώς θα ορίσει το τι είναι βιοπληροφορική. Κάποιοι επιχειρούν να εντοπίσουν τις σχετικές δημοσιεύσεις με χρήση κάποιων καθορισμένων εκ των προτέρων λέξεων-κλειδιών (keywords ή MESH terms), άλλοι επιλέγουν από την αρχή τα επιστημονικά περιοδικά που θεωρούν ότι είναι χαρακτηριστικά του χώρου, ενώ άλλοι ακολουθουν μια μικτή στρατηγική. Σε κάθε περίπτωση, τα αποτελέσματα είναι ιδιαίτερα ενδιαφέροντα και κάποια από αυτά θα προσπαθήσουμε να περιγράψουμε παρακάτω.

Σε μια από τις πρώτες τέτοιες μελέτες, οι Patra and Mishra (Patra & Mishra, 2006) χρησιμοποίησαν κάποια γενικά MESH terms όπως "Bioinformatics" OR "Bioinformatics" OR "Computational Biology" OR "Computational Molecular Biology" OR "Biology Computational" OR "Molecular Biology; Computational" OR "Genomics" για να πραγματοποιήσουν αναζήτηση στην PUBMED και συγκέντρωσαν 16.178 επιστημονικά άρθρα, τα οποία είχαν δημοσιευθεί μέχρι το 2004, προερχόμενα από 1806 διαφορετικά επιστημονικά περιοδικά. Όπως θα ανέμενε κανείς, η αύξηση την περίοδο πριν από το 2000 ήταν εκθετική. Ενδεικτικά, το 1990 είχαν δημοσιευθεί μόνο 12 άρθρα, ενώ το 2000 ο αντίστοιχος αριθμός ξεπέρασε τα 1000.

Η μεγάλη πλειοψηφία των εργασιών αυτών ήταν άρθρα σε περιοδικά (98%) και ήταν γραμμένες στα Αγγλικά (97%). Οι επιστήμονες από τις ΗΠΑ είχαν το μεγαλύτερο μερίδιο στους συγγραφείς (42%) ακολουθούμενοι από τους Βρετανούς (10%), τους Γερμανούς (6%) και τους Ιάπωνες (4%).

Bioinformatics
Nucleic Acids Research
Genome Research
Science
Nature
Proceedings of the National Academy of Sciences USA
Proteomics
Genome Biology
Journal of Molecular Biology
Proteins
Nature Biotechnology
BMC Bioinformatics
Pacific Symposium on Biocomputing
Journal of Computational Biology
Tanpakushitsu Kakusan Koso
Journal of Biological Chemistry
Drug Discovery Today
Trends in Biotechnology
Genomics
Briefing in Bioinformatics

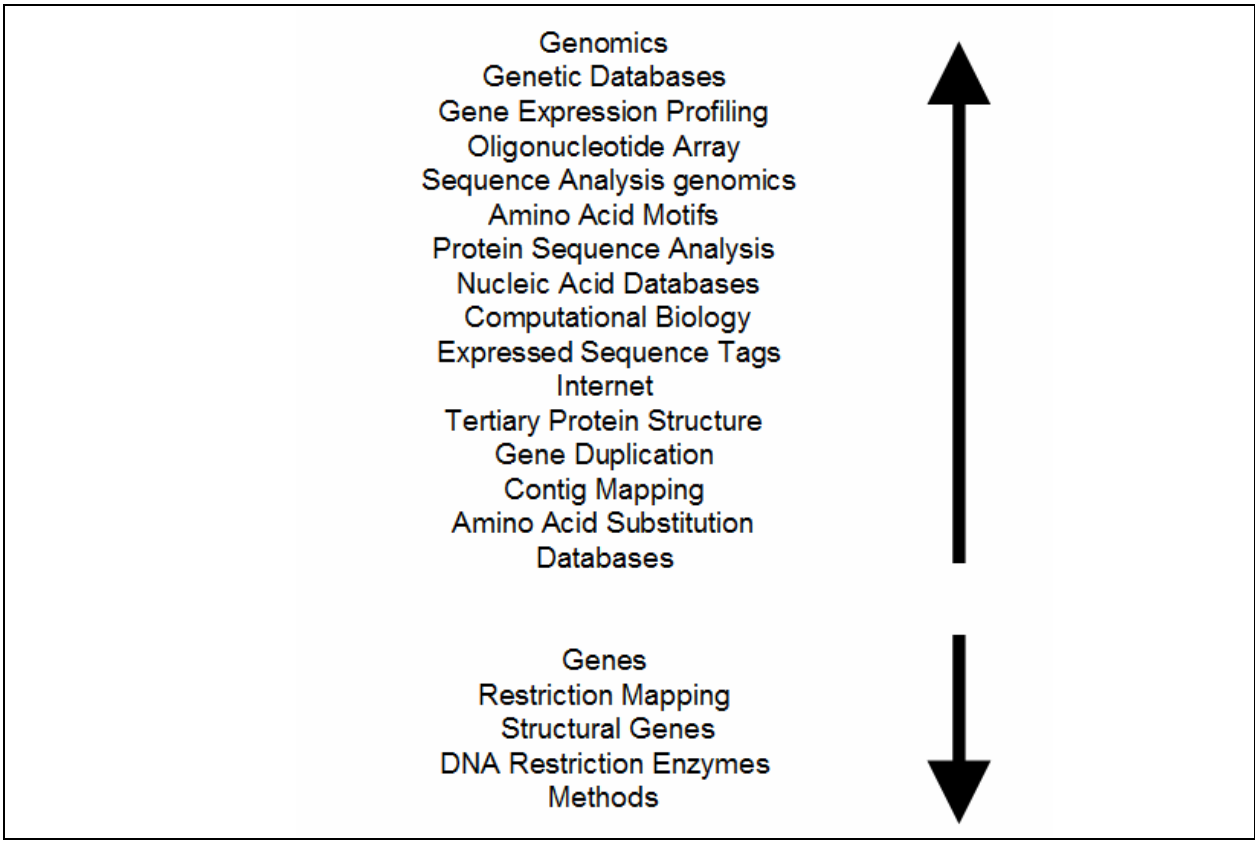
Πίνακας 1.1: *Τα 20 κύρια περιοδικά βιοπληροφορικής που εντοπίστηκαν στη μελέτη των (Patra & Mishra, 2006)*

Από την κατανομή των άρθρων ανά περιοδικό, εντόπισαν τα 20 «κύρια» επιστημονικά περιοδικά στα οποία είχαν δημοσιευθεί το 1/3 των εργασιών αυτών. Τα περιοδικά αυτά φαίνονται στον Πίνακα 1.1. Όπως είναι αναμενόμενο, κάποια από αυτά τα περιοδικά είναι αποκλειστικά περιοδικά βιοπληροφορικής (Bioinjormatics, BMC Bioinformatics, Pacific Symposium on Biocomputing, Journal of Computational Biology κ.ο.κ.), αλλά υπάρχουν και πολλά από τα κορυφαία περιοδικά βιολογικού (Nucleic Acids Research, Genome Research, Genome Biology, Journal of Molecular Biology) ή γενικότερου ενδιαφέροντος (Science, Nature, Proceedings of the National Academy of Sciences USA). Παρατηρήθηκε επίσης, ότι τα περισσότερα από τα περιοδικά αυτά είχαν αυξήσει το Impact Factor τους από το 2001 και μετά, πιθανότατα λόγω αυξημένων αναφορών από τις εργασίες βιοπληροφορικής, ενώ κάποια από αυτά είχαν αλλάξει ακόμα και το όνομά τους για να ανταποκριθούν καλύτερα στις νέες συνθήκες (π.χ. το Computer Application in Biosciences μετονομάστηκε σε Bioinfonnatics το 1998, ενώ το PCR Methods and Its Applications μετονομάστηκε σε Genome Research το 1994).

Όσον αφορά τον αριθμό των συγγραφέων σε κάθε άρθρο, η μελέτη βρήκε ότι το 23% των εργασιών είχε γραφτεί από έναν συγγραφέα, ενώ το 23% από δύο. Ο μέγιστος αριθμός συγγραφέων σε ένα άρθρο ήταν 40 και βρέθηκαν 67 τέτοια άρθρα. Τα περισσότερα από αυτά τα άρθρα εμφανίστηκαν μετά το 2000 κάτι που προφανώς έχει σχέση με τα τεράστια συνεργατικά consortia που ασχολήθηκαν με την αλληλούχιση γονιδιωμάτων. Συνολικά, υπήρχαν 39.435 συγγραφείς για τα 16.178 άρθρα (2,43 συγγραφείς/εργασία). Παρόλα αυτά, το 73,58% των συγγραφέων είχε συνεισφορά μόνο σε ένα άρθρο, ενώ το 14,34% σε δύο και το 5,30% σε τρία. Τα αποτελέσματα αυτά είναι σύμφωνα με το νόμο του Lotka που λέει ότι ο αριθμός των συγγραφέων που έχει n εργασίες, είναι περίπου ίσος με το $1/n^2$ αυτών που έχουν μόνο μία. Σύμφωνα με αυτόν το νόμο μόνο το 6% των συγγραφέων σε ένα ερευνητικό πεδίο θα έχει πάνω από 10 εργασίες. Τα αποτελέσματα αυτά εξηγούνται αν αναλογιστούμε ότι πολλοί συγγραφείς κάνουν μία ή δύο εργασίες και μετά εγκαταλείπουν την έρευνα σε αυτό το αντικείμενο, είτε γιατί είναι φοιτητές που δεν συνεχίζουν την ερευνητική καριέρα, είτε γιατί αλλάζουν αντικείμενο. Επίσης, η διεπιστημονικότητα της βιοπληροφορικής συντείνει στο να υπάρχουν και αρκετοί καταξιωμένοι επιστήμονες από άλλους χώρους (μαθηματικά, πληροφορική, βιολογία), οι οποίοι μόνο περιστασιακά ενεπλάκησαν στη συγγραφή εργασιών βιοπληροφορικής.

Σε μια άλλη εργασία του 2006, οι Perez-Iratxeta, Andrade-Navarro και Wren (Perez-Iratxeta, Andrade-Navarro, & Wren, 2007) έκαναν μια ανάλυση των λέξεων σε όλα τα άρθρα της PUBMED με σκοπό να εντοπίσουν βιοπληροφορικές εργασίες εστιάζοντας σε 3 διαφορετικούς τομείς (υπολογισμούς, διαδίκτυο

και βάσεις δεδομένων). Έκαναν την αναζήτηση μεταξύ των ετών 1996 και 2005 αναζητώντας όρους (MESH terms) που παραπέμπουν σε υπολογιστική ανάλυση βιολογικών δεδομένων όπως: 'comput*', '*informatic*', 'algorithm*', 'software' ή 'database' ενώ επιπλέον αναζήτηση έγινε για λέξεις κλειδιά όπως 'internet', 'online', 'world wide web', 'web-based', 'http:*' και 'ftp:*'.



Εικόνα 1.11: *Η εξέλιξη των όρων όπως αποτυπώθηκε στη μελέτη των (Perez-Iratxeta, Andrade-Navarro, & Wren, 2007)*

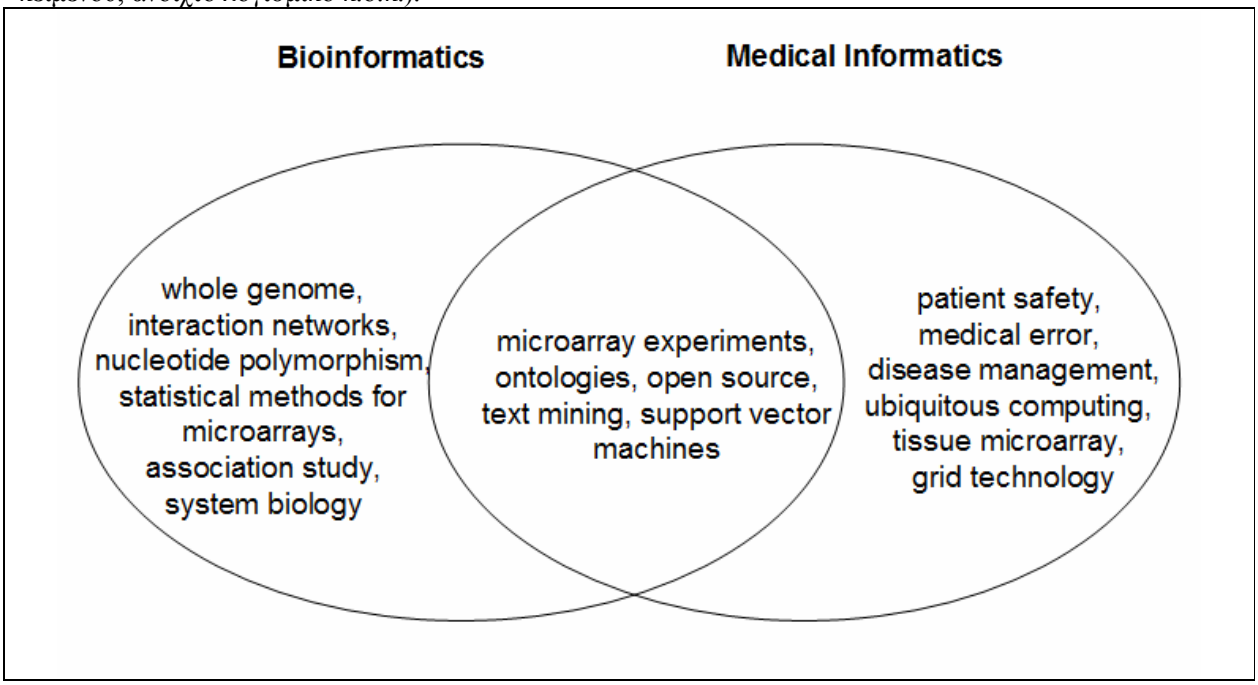
Η μελέτη αυτή έδειξε ότι το ποσοστό των εργασιών που χρησιμοποιούν υπολογιστικές τεχνικές στη βιοϊατρική έρευνα αυξήθηκε από το 1,6% το 1975, στο 10% το 2005. Η χρήση διαδικτυακών πηγών αυξήθηκε από το 0,05% στο 0.87% την ίδια περίοδο με τη μεγαλύτερη άνοδο να συμβαίνει τη δεκαετία του 1990 με την εξάπλωση του διαδικτύου και των H/Y. Επίσης, παρόμοια αύξηση υπάρχει και στην αναφερόμενη χρήση βάσεων δεδομένων, κάτι που απεικονίζει και την τεράστια αύξηση των δεδομένων που βρίσκονται κατετεθειμένες σε αυτές αλλά και την αντίστοιχη τεχνολογική πρόοδο που έκανε εύκολη τη συλλογή και αποθήκευση αυτών των δεδομένων. Γενικά, τα αποτελέσματα έδειξαν ότι η αύξηση στη χρήση υπολογιστικών μεθόδων συνέβη πρώτα, μετά ακολούθησε η εξάπλωση του διαδικτύου και στο τέλος έγινε η εμφάνιση των βάσεων δεδομένων, λόγω την ανάπτυξης των άλλων δύο τεχνικών. Τέλος, η χρονική ανάλυση των MESH terms υπέδειξε μια μεγάλη ομάδα όρων για τις οποίες υπήρξε μια τεράστια αύξηση στην εμφάνιση τους τα χρόνια 2000-2003 σε σχέση με τα προηγούμενα, και μια αντίστοιχη ομάδα όρων με μειωμένη χρήση (Εικόνα 1.11). Οι όροι αυτοί αντικατοπτρίζουν την μετάβαση από τις απλές μοριακές και βιοχημικές τεχνικές (π.χ. τη μελέτη ενός γονιδίου) στις υπολογιστικές αναλύσεις γονιδιωμάτων και μεγάλων συνόλων δεδομένων (κάτι που είναι χαρακτηριστικό της βιοπληροφορικής).

Μια άλλη μελέτη από το 2007, η οποία προέκυψε από μια μεγάλη διεπιστημονική συνεργασία (SYMBIOMATICS), αποσκοπούσε στο να μετρήσει με μια κλασική μέθοδο ανάλυσης κειμένου (bigrams, δηλαδή με τα ζευγάρια λέξεων) τη διαχρονική πορεία της σχετικής βιλιογραφίας, να συγκρίνει τις «πρόσφατες» (2000-2005) με τις «παλιές» (1990-1990) εργασίες και να εντοπίσει έτσι τις αναδυόμενες τάσεις στο ερευνητικό πεδίο της βιοπληροφορικής, αλλά και επιπλέον, να εντοπίσει τις περιοχές σύγκλισης και διαφοροποίησης της βιοπληροφορικής με την ιατρική πληροφορική (Rebholz-Schuhman et al., 2007). Τα περιοδικά που θεωρήθηκαν από τους συγγραφείς ότι ανήκουν στις δύο κατηγορίες δίνονται στον Πίνακα

Βιοπληροφορική	Ιατρική Πληροφορική
Bioinformatics	AMIA Annu Symp Proc
Biosystems	Artif Intell Med
BMC Bioinformatics	BMC Med Inform Decis Mak
Brief Bioinform	Int J Med Inform
Comput Methods Programs Biomed	J Am Med Inform Assoc
IEEE Trans Inf Technol Biomed	Medinfo
J Bioinform Comput Biol	Methods Inf Med
J Biomed Inform	Proc AMIA Symp
J Comput Aided Mol Des	
J Comput Biol	
Pac Symp Biocomput	

Πίνακας 1.2: *Τα περιοδικά βιοπληροφορικής και ιατρικής πληροφορικής που χρησιμοποιήθηκαν στη μελέτη SYMBIOMATICS*

Η ανάλυση έδειξε καταρχάς ότι η μεγάλη αύξηση της βιβλιογραφίας της βιοπληροφορικής έγινε την περίοδο 2000-2005, ενώ η αντίστοιχη αύξηση της βιβλιογραφίας της ιατρικής πληροφορικής είχε συντελεστεί την δεκαετία 1990-2000. Για όλη τη δεκαπενταετία της ανάλυσης (1990-2005) οι πιο συνηθισμένες λέξεις κλειδιά για τη Βιοπληροφορική ήταν «gene expression», «amino acid», και «protein sequence», ενώ για την ιατρική πληροφορική «information system», «health care» και «decision support». Ιδιαίτερο ενδιαφέρον όμως προκύπτει από τη διαχρονική ανάλυση και τη σύγκριση της περιόδου 1990-2000 με την αντίστοιχη περίοδο 2000-2005 από την οποία προκύπτουν οι αναδυόμενες τάσεις στα δύο πεδία καθώς και οι περιοχές σύγκλισής τους. Όπως φαίνεται στην Εικόνα 1.12 έννοιες που σχετίζονται με τις μικροσυστοιχίες, με τις οντολογίες, το ανοικτό λογισμικό, την ανάλυση κειμένου και τα Support Vector Machines είναι κοινές και στους δύο κλάδους και καταδεικνύουν μια μεγάλη περιοχή διεπαφής και συνέργιας. Οι συνέργιες αυτές των δύο περιοχών υπάρχουν και αυξάνονται διαχρονικά για μια σειρά από λόγους: Πρώτον, και οι δύο επιστημονικές περιοχές επωφελούνται από, και ασχολούνται με, τις νέες τεχνολογίες της βιοϊατρικής (π.χ. μικροσυστοιχίες), έστω και αν τις αντιμετωπίζουν με διαφορετικό τρόπο (ανάπτυξη μεθόδων από τη βιοπληροφορική, εφαρμογή στην κλινική πράξη από την ιατρική πληροφορική). Δεύτερον, και οι δύο επωφελούνται από νέες ανακαλύψεις στα μαθηματικά και την πληροφορική (π.χ. support vector machines). Τέλος, υπάρχει οριζόντια διάχυση καθώς τεχνολογίες και μεθοδολογίες που αναπτύχθηκαν αρχικά στον ένα κλάδο διαχέονται τελικά και στον άλλον και κατόπιν αναπτύσσονται παράλληλα με όφελος και για τους δύο (π.χ. οντολογίες, ανάλυση κειμένου, ανοιχτό λογισμικό κ.ο.κ.).



Εικόνα 1.12: *Τα ερευνητικά πεδία της βιοπληροφορικής και ιατρικής πληροφορικής με τις ομοιότητες και τις διαφορές τους, όπως βρέθηκαν από τη μελέτη SYMBIOMATICS*

Τέλος, το 2014 έγινε η πιο πρόσφατη βιβλιομετρική εργασία ειδικά για το πεδίο της βιοπληροφορικής (Song, Kim, Zhang, Ding, & Chambers, 2014). Οι ερευνητές προσπάθησαν να αντιμετωπίσουν τους περιορισμούς που είχαν οι προηγούμενες αναλύσεις τόσο από πλευράς κάλυψης της βιβλιογραφίας, όσο και από την πλευρά της μεθοδολογίας της ανάλυσης, ενώ φυσικά, το γεγονός ότι μιλάμε για μια σύγχρονη εργασία βοηθάει ιδιαίτερα στο να κατανοήσουμε τις τελευταίες εξελίξεις στον τομέα. Συνδυάζοντας δεδομένα από προηγούμενες μελέτες, οι ερευνητές κατέληξαν σε ένα μεγάλο σύνολο περιοδικών (μεγαλύτερο από τις προηγούμενες μελέτες), από τα οποία μόνο το 73% καλύπτεται από τη βάση δεδομένων του WoS (αλλά όλα είναι καταχωρημένα στην PUBMED). Με την επιλογή να αναλύσουν τα πλήρη κείμενα αντί για τις περιλήψεις, μπόρεσαν να κάνουν πιο ολοκληρωμένη ανάλυση κειμένου, με αντίτιμο το γεγονός ότι τα πλήρη κείμενα ήταν κατατεθειμένα στην PubmedCentral, το υποσύνολο της PUBMED στο οποίο γίνεται κατάθεση των εργασιών που δημοσιεύονται με το πρότυπο ανοιχτής πρόσβασης (open access) ή κατατίθενται από άλλα περιοδικά αλλά αφού έχει περάσει κάποιος χρόνος από την αρχική δημοσίευση. Με αυτόν τον τρόπο, κάποια περιοδικά ή ακόμα και κάποια άρθρα κάποιων περιοδικών δεν συμπεριλήφθηκαν στην ανάλυση. Ένα άλλο πλεονέκτημα αυτής της εργασίας ήταν ότι μπόρεσαν ταυτόχρονα να αναλύσουν και τις αναφορές των εργασιών, με αποτέλεσμα να εξαχθούν συμπεράσματα για τα πιο επιδραστικά περιοδικά, τους συγγραφείς, τις εργασίες αλλά και τα ιδρύματα.

Τα περιοδικά που χρησιμοποιήθηκαν δίνονται στον Πίνακα 1.3 και η γενικότερη ανάλυση δείχνει ότι η αύξηση της σχετικής βιβλιογραφίας είναι εκθετική, κάτι που επιβεβαιώνει τις προηγούμενες μελέτες. Κατά την περίοδο 2000-2003, το κυρίαρχο αντικείμενο ήταν η μελέτη των πρωτεϊνών και κυρίως η λειτουργική μελέτη τους. Κατά την περίοδο 2004-2007 τα αντικείμενα γίνονται πιο ετερογενή και περιλαμβάνουν τη δομική ανάλυση γονιδίων, τον εγκέφαλο, τον καρκίνο και τους ιούς. Κατά την περίοδο 2008-2011, τα αντικείμενα συνεχίζουν να διαφοροποιούνται αλλά παρουσιάζουν ομοιότητες με την δεύτερη περίοδο. Παρόλα αυτά, νέοι όροι εμφανίζονται σε αυτή την περίοδο όπως mutation και RNA. Γενικά, τα πιο συνηθισμένα αντικείμενα όλης της περιόδου περιλαμβάνουν όρους όπως protein binding, algorithm/method, cell/model, network/interaction, genome sequence, immune/virus, gene expression, genetic/evolution, database/software, gene transcription, DNA/chromosome, ontology/mining, gene/genomics και cancer/cell.

Σχετικά με τις χώρες προέλευσης των εργασιών οι ΗΠΑ, η Μεγάλη Βρετανία και η Γερμανία βρίσκονται σταθερά στις πρώτες θέσεις σε όλες τις περιόδους, ακολουθούμενες από τη Γαλλία και τον Καναδά. Σταθερή άνοδο παρουσιάζουν η Κίνα και η Ιαπωνία. Η πρώτη χώρα, από μια θέση μεγαλύτερη της 20^{ης} την περίοδο 2000-2003, φτάνει στην περίοδο 2009-2011 να βρίσκεται στην 6^{η} θέση, ενώ η δεύτερη από την 9^{η} θέση ανεβαίνει σταδιακά στην 7^{η}. Η Ιταλία και η Ισπανία βρίσκονται σταθερά μέσα στην πρώτη δεκάδα ενώ η Ελλάδα όπως ίσως θα αναμέναμε, λόγω μεγέθους αλλά και ΑΕΠ, βρίσκεται σε θέση μεγαλύτερη της 20^{ης} σε όλες τις περιόδους (περισσότερα για την Ελλάδα στην επόμενη ενότητα). Τα πρότυπα της σχετικής συνεισφοράς και της κατάταξης των διαφόρων κρατών στην έρευνα στη βιοπληροφορική, ακολουθούν τα γενικότερα πρότυπα που έχουν βρεθεί για το σύνολο της επιστημονικής παραγωγικότητας, τόσο από βιβλιομετρικές επιστημονικές μελέτες (King, 2004), όσο και από κατατάξεις της παγκόσμιας βιβλιογραφίας (http://www.natureindex.com/).

BMC Bioinformatics	Source Code for Biology and Medicine
BMC Genomics	Advanced Bioinformatics
PLoS Biology	BioData Mining
Genome Biology	Journal of Computational Neuroscience
PLoS Genetics	Journal of Proteome Research
PLoS Computational Biology	Journal of Biomedical Semantics
BMC Research Notes	Journal of Computer-Aided Molecular Design
Bioinformatics	Genome Integration
Molecular Systems Biology	Journal of Molecular Modeling
BMC Systems Biology	Bulletin of Mathematical Biology
Comparative and Functional Genomics	Pharmacogenetics and Genomics
Bioinformation	Statistical Methods in Medical Research
Theoretical Biology and Medical Modeling	Neuroinformatics
Human Molecular Genetics	Genomics

The EMBO Journal	Protein Science
Cancer Informatics	Physiological Genomics
Genome Medicine	Trends in Genetics
Evolutionary Bioinformatics	Journal of Proteomics
Biochemistry	Proteomics
Algorithms for Molecular Biology	Trends in Biochemical Sciences
EURASIP Journal on Bioinformatics and Systems Biology	Journal of Biotechnology
Journal of Molecular Biology	Trends in Biotechnology
Molecular & Cellular Proteomics	Briefings in Functional Genomics & Proteomics
Mammalian Genome	Journal of Theoretical Biology

Πίνακας 1.3: *Τα περιοδικά βιοπληροφορικής που μελετήθηκαν στην εργασία των (Song, Kim, Zhang, Ding, & Chambers, 2014).*

Όσον αφορά την κατάταξη των πανεπιστημίων το Stanford University, το Harvard University και το University of Washington βρίσκονται σταθερά ψηλά στη σχετική λίστα. Το Stanford ήταν 3^{ο} την περίοδο 2000-2003 και 1^{ο} από το 2004 και μετά. Το Harvard ήταν 6^{ο}, 2^{ο} και 3^{ο} αντίστοιχα στις περιόδους 2000-2003, 2004-2007 και 2009-2011. Τέλος, το University of Washington ήταν 5^{ο} στις δύο πρώτες περιόδους και 2^{ο} στην τελευταία. Τρία πανεπιστήμια είχαν σταθερά ανοδική πορεία στις αντίστοιχες περιόδους, το University of Cambridge (11^{ο}, 8^{ο} και 5^{ο}), το University College London (11^{ο}, 11^{ο} και 10^{ο}), ενώ το University of Oxford δεν ήταν καν στη λίστα με τα κορυφαία ιδρύματα την περίοδο 2000-2003, αλλά ανέβηκε στην 10^{η} θέση την περίοδο 2004-2008 και στην 6^{η} την περίοδο 2009-2011. Από την άλλη μεριά, το Brandeis University που ήταν 1^{ο} την περίοδο 2000-2003, έπεσε στην 12^{η} θέση την περίοδο 2004-2007 και δεν συμπεριλαμβάνεται καν στη λίστα την περίοδο 2008-2011. Παρόμοια πτωτική πορεία είχε το University of California, Berkeley το οποίο έπεσε από τη 2^{η} θέση στην 7^{η} και τελικά στην 14^{η} στις αντίστοιχες περιόδους.

Η ανάλυση έδωσε επίσης δεδομένα για τις πιο επιδραστικές εργασίες στο χώρο καθώς και πληροφορίες για τους συγγραφείς που συμμετείχαν σε αυτές. Η ανάλυση αυτή είναι ιδιαίτερα χρήσιμη γιατί έτσι μπορούν να αναγνωριστούν κάποια από τα ερευνητικά πεδία που κυριάρχησαν στις επόμενες περιόδους. Για την περίοδο 2000-2003 η εργασία με τις περσσότερες αναφορές είχε τίτλο "Gene ontology: tool for the unification of biology" και δημοσιεύτηκε στο Nature Genetics με συγγραφείς το Gene Ontology Consortium το οποίο αποτελούταν από 20 ερευνητές. Οκτώ από αυτούς συγκαταλέγονται στους πιο επιδραστικούς συγγραφείς αυτής της χρονικής περιόδου (D. Botstein, G. Rubin, G. Sherlock, M. Ashburner, J. Cherry, C. Ball, J. Matese, H. Butler). Η δεύτερη σε αριθμό αναφορών εργασία, ήταν η δημοσίευση του ανθρώπινου γονιδιώματος ("Initial sequencing and analysis of the human genome") στο Nature. Οι συγγραφείς ήταν 249 από 48 διαφορετικά ινστιτούτα. Η 3^{η} πιο σημαντική εργασία αυτής της περιόδου ήταν το "Significance analysis of microarrays applied to the ionizing radiation response" από τους V. Tusher, R. Tibshirani, και G. Chu, από το Stanford. Ο R. Tibshirani επίσης εμφανίζεται στη 12^{η} θέση των συγγραφέων την ίδια περίοδο. Κατά την περίοδο 2004-2007, η εργασία με τις περισσότερες αναφορές είχε τίτλο "Bioconductor: open software development for computational biology and bioinformatics" και την συνέγραψαν 25 συγγραφείς από 19 ινστιτούτα. Ανάμεσα τους 4 βρίσκονται στη λίστα με τους πιο επιδραστικούς επιστήμονες της ίδιας περιόδου. Η δεύτερη εργασία της περιόδου αυτής ήταν η εργασία που περιέγραφε το στατιστικό πακέτο R, "R: A language and environment for statistical computing" και η 3^{η} είχε τίτλο "Transcriptional regulatory code of a eukaryotic genome" από 20 συγγραφείς από 4 διαφορετικά ινστιτούτα. Τέλος, κατά την περίοδο 2008-2011, η εργασία με τις περισσότερες αναφορές ήταν η εργασία που παρουσίαζε τη βάση δεδομένων PFAM "The Pfam protein families database" με 13 συγγραφείς από 3 διαφορετικά ινστιτούτα (ανάμεσά τους ο A Bateman και ο R. Durbin, οι οποίοι βρίσκονται σταθερά μέσα στη λίστα των πιο επιδραστικών επιστημόνων του χώρου). Η δεύτερη στη σειρά εργασία αφορούσε την περιγραφή της KEGG, "KEGG for linking genomes to life and the environment" με 11 συγγραφείς από 3 διαφορετικά ινστιτούτα της Ιαπωνίας, ενώ η τρίτη είχε τίτλο "Mapping short DNA sequencing reads and calling variants using mapping quality scores" με συγγραφείς τους H. Li, J Ruan και R. Durbin. Αυτά τα δεδομένα δείχνουν την κατεύθυνση της έρευνας την τελευταία 10ετία και βρίσκονται σε συμφωνία με τα δεδομένα που αναφέρθηκαν στις προηγούμενες μελέτες (ανάπτυξη των μεθόδων αλληλούχισης, ανάπτυξη στατιστικών μεθόδων ελέγχου της γονιδιακής έκφρασης, ανάπτυξη του ανοιχτού λογισμικού βιοπληροφορικής αλλά και των βάσεων δεδομένων και των οντολογιών).

Από την ανάλυση των αναφορών προέκυψε επίσης και η κατάταξη των επιδραστικών περιοδικών του χώρου της βιοπληροφορικής, η οποία βέβαια αναμένουμε να έχει παρόμοια δομή με την αντίστοιχη κατάταξη

με βάση το Impact Factor. Έτσι, βλέπουμε ότι σε όλες τις περιόδους τα περιοδικά Proceedings of the National Academy of Sciences, Nucleic Acids Research, Nature, Bioinformatics και Science βρίσκονται σταθερά στην κορυφαία πεντάδα των περιοδικών. Το BMC Bioinformatics το οποίο έκανε την εμφάνιση του σχετικά αργότερα είναι 6^{ο} την περίοδο 2004-2007 και 5^{ο} την περίοδο 2008-2011. Εκτός από το BMC Bioinformatics, την περίοδο 2004-2007 εμφανίζεται και το PLoS Biology, το BMC Genomics, και το Nature Reviews Genetics με τη σειρά κατάταξής τους να αυξάνει συνεχώς. Νέα περιοδικά που έκαναν δυναμική εμφάνιση στα 20 καλύτερα την περίοδο 2008-2009 ήταν το PLoS One, PLoS Genetics, PLoS Computational Biology, Nature Biotechnology και το Nature Methods. Αξίζει να σημειωθεί εδώ ότι τα περισσότερα από τα κορυφαία περιοδικά (με εξαίρεση τα Bioinformatics, BMC Bioinformatics και PLoS Computational Biology) είναι βιολογικά περιοδικά ή περιοδικά γενικότερου ενδιαφέροντος στα οποία δημοσιεύονται και εργασίες βιοπληροφορικής. Αυτό αφενός κάνει δύσκολη την αναζήτηση της σχετικής βιβλιογραφίας για παρόμοιες αναλύσεις, αλλά παράλληλα δείχνει και τη σπουδαιότητα που έχει η βιοπληροφορική στο πλαίσιο της σύγχρονης βιολογικής και βιοϊατρικής έρευνας.

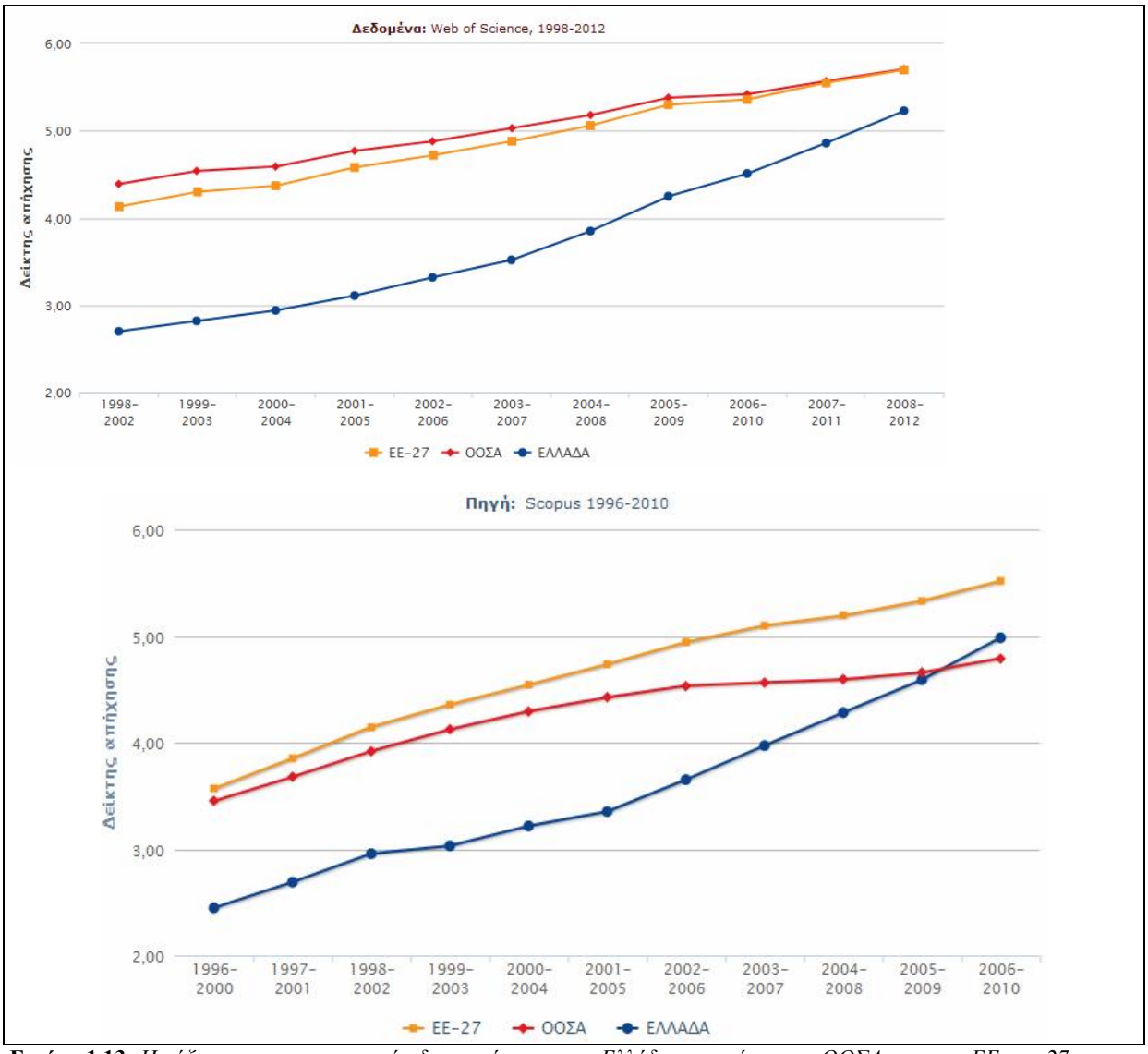
Συνολικά, από τη μελέτη αυτή προέκυψε ότι το πεδίο της βιοπληροφορικής έχει υποστεί σημαντικές αλλαγές κατά την τελευταία 10ετία και εξελίσσεται παράλληλα με άλλες ειδικότητες της βιοϊατρικής, καθώς η εστίαση και το αντικείμενο της μελέτης αλλάζει με την πάροδο του χρόνου (γονιδιακή έκφραση, γονιδιώματα, πολυμορφισμοί, πολύπλοκα συστήματα κ.ο.κ.). Η ανάπτυξη των υπολογιστικών προσεγγίσεων έχει βοηθήσει την εξάπλωση των βιολογικών βάσεων δεδομένων και των εργαλείων ανάλυσής τους ακόμα και σε άλλες συγγενικές ειδικότητες. Ειδικότερα, οι υπολογιστικές τεχνικές έγιναν αναπόσπαστο τμήμα της βιοϊατρικής έρευνας την περίοδο 2000-2003, ενώ μετά το 2004 αυξήθηκε και η ανάγκη δημιουργίας και διαχείρισης βάσεων βιολογικών δεδομένων. Τέλος, σημαντικά εργαλεία και μεθοδολογίες της βιοπληροφορικής που αναπτύχθηκαν αυτή την περίοδο, όπως οι μικροσυστοιχίες οι οντολογίες και οι μεθοδολογίες ανάλυσης βιολογικών δικτύων έγιναν βασικά κομμάτια της σύγχρονης βιοπληροφορικής έρευνας αλλά όπως είδαμε και προηγουμένως, οι μικροσυστοιχίες και οι οντολογίες έγιναν και βασικά κομμάτια της ιατρικής πληροφορικής.

1.4. Η κατάσταση στην Ελλάδα

Στην ενότητα αυτή θα μελετήσουμε την κατάσταση της βιοπληροφορικής στην Ελλάδα. Θα δούμε τις επιστημονικές δημοσιεύσεις που έχουν προέλθει από Ελληνικά ιδρύματα, την κατάσταση της επιστημονικής κοινότητας της βιοπληροφορικής στην Ελλάδα, αλλά και την εκπαίδευση.

1.4.1. Η έρευνα στην Ελλάδα

Γενικά, παρά τις περί του αντιθέτου κραυγές που ακούγονται καθημερινά στα ΜΜΕ, η επιστημονική έρευνα στην Ελλάδα πάει καλά (ειδικά αν αναλογιστούμε τα χρήματα που δαπανώνται σε αυτή, βλ. παρακάτω). Όλες οι αποτιμήσεις της ερευνητικής δραστηριότητας που έχουν γίνει τα τελευταία χρόνια από το Εθνικό Κέντρο Τεκμηρίωσης (ΕΚΤ), δείχνουν ότι τόσο ο αριθμός των δημοσιεύσεων όσο και των αναφορών που παίρνουν εργασίες προερχόμενες από ελληνικά ιδρύματα αυξάνονται συνεχώς, αλλά το πιο σημαντικό είναι ότι αυξάνονται και οι ποιοτικοί δείκτες με αποτέλεσμα η Ελλάδα να συγκλίνει σταδιακά προς το μέσο όρο των χωρών της ΕΕ και του ΟΟΣΑ (Εικόνα 1.13) σε μέτρα που αφορούν την ποιότητα (π.χ. το μέσο αριθμό αναφορών ανά εργασία). Όλες οι μελέτες του ΕΚΤ την τελευταία δεκαετία συγκλίνουν στο αποτέλεσμα αυτό, ανεξαρτήτως της βάση δεδομένων βιβλιογραφίας που χρησιμοποιείται (Σαχίνη, Μάλλιου, & Χούσος, 2012; Σαχίνη, Μάλλιου, Χούσος, & Καραϊσκος, 2013; Σαχίνη, Μάλλιου, Χούσος, & Καραΐσκος, 2014) και το ίδιο φαίνεται να ισχύει ειδικά και για τον τομέα των βιοϊατρικών επιστημών (Σαχίνη, Μάλλιου, Χούσος, & Καραϊσκος, 2012). Οι ίδιες μελέτες δείχνουν καθαρά ότι το μεγαλύτερο σε όγκο μέρος της ερευνητικής δραστηριότητας της χώρας προέρχεται από τα Ελληνικά Πανεπιστήμια, αν και φυσικά υπάρχουν και ιδιαίτερα αξιόλογα Ερευνητικά Κέντρα, αλλά και νησίδες αριστείας στα ΤΕΙ.



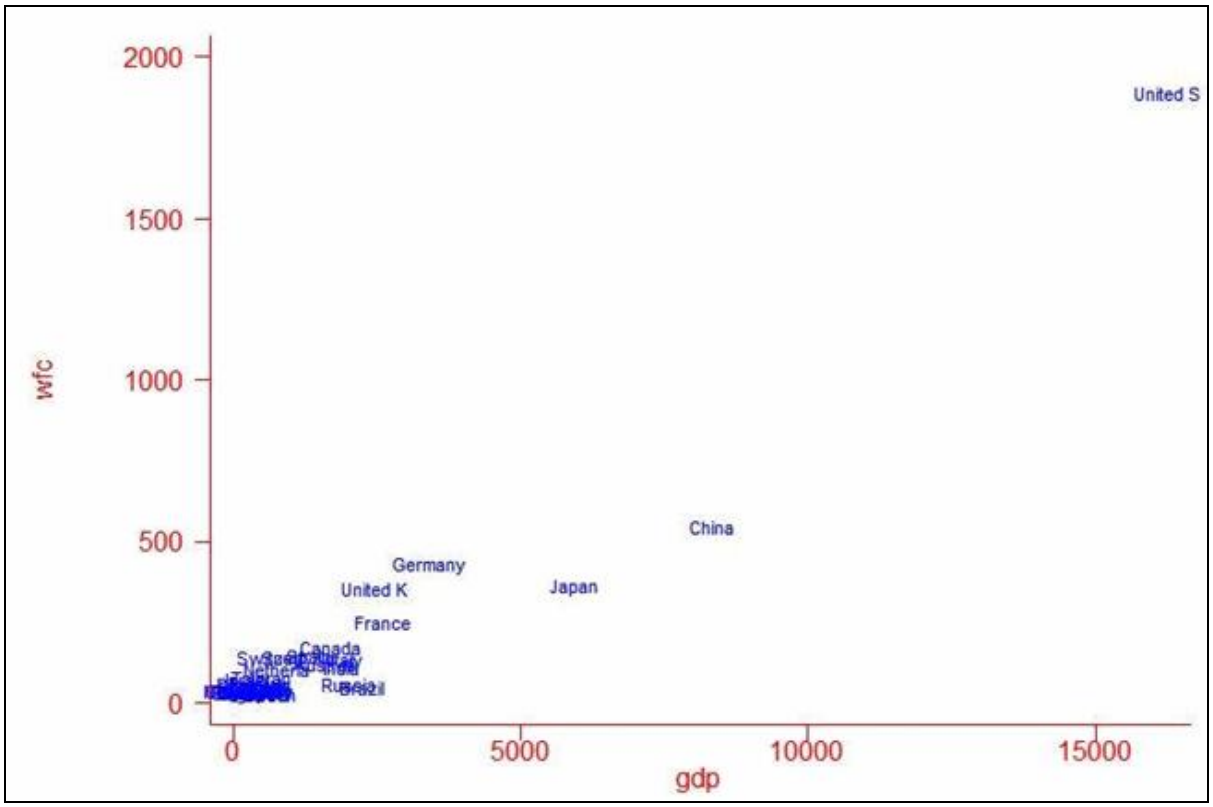
Εικόνα 1.13: *Η αύξηση των επιστημονικών δημοσιεύσεων στην Ελλάδα, στις χώρες του ΟΟΣΑ και στην ΕΕ των 27*

Φυσικά, πάντα χρειάζεται προσοχή στο πώς σταθμίζουμε τέτοιες αναλύσεις γιατί θα πρέπει να λαμβάνουμε υπόψη μας και παράγοντες όπως ο πληθυσμός μιας χώρας αλλά και το ΑΕΠ (Ακαθάριστο Εγχώριο Προϊόν) αυτής. Για παράδειγμα, το επιστημονικό περιοδικό Nature, δημοσίευσε πρόσφατα μια μελέτη (http://www.natureindex.com) που απαριθμεί τα άρθρα υψηλής επιστημονικής απήχησης του τελευταίου χρόνου (τις εργασίες που έχουνε δημοσιευθεί σε μια συλλογή απο τα σημαντικότερα διεθνή επιστημονικά περιοδικά). Με βάση αυτούς τους πίνακες (Εικόνα 1.14), που μετράνε παραγωγικότητα έρευνας υψηλού επιπέδου χωρίς καμία άλλη στάθμιση (δείκτης WFC), η Ελλάδα, κατατάσεται στην 32η θέση στον κόσμο, ενώ βρίσκεται στην προτελευταία ανάμεσα στις χώρες της Ευρωπαϊκής Ένωσης (ΕΕ) πριν τη διεύρυνση του 2004 (προσπερνά το Λουξεμβούργο).

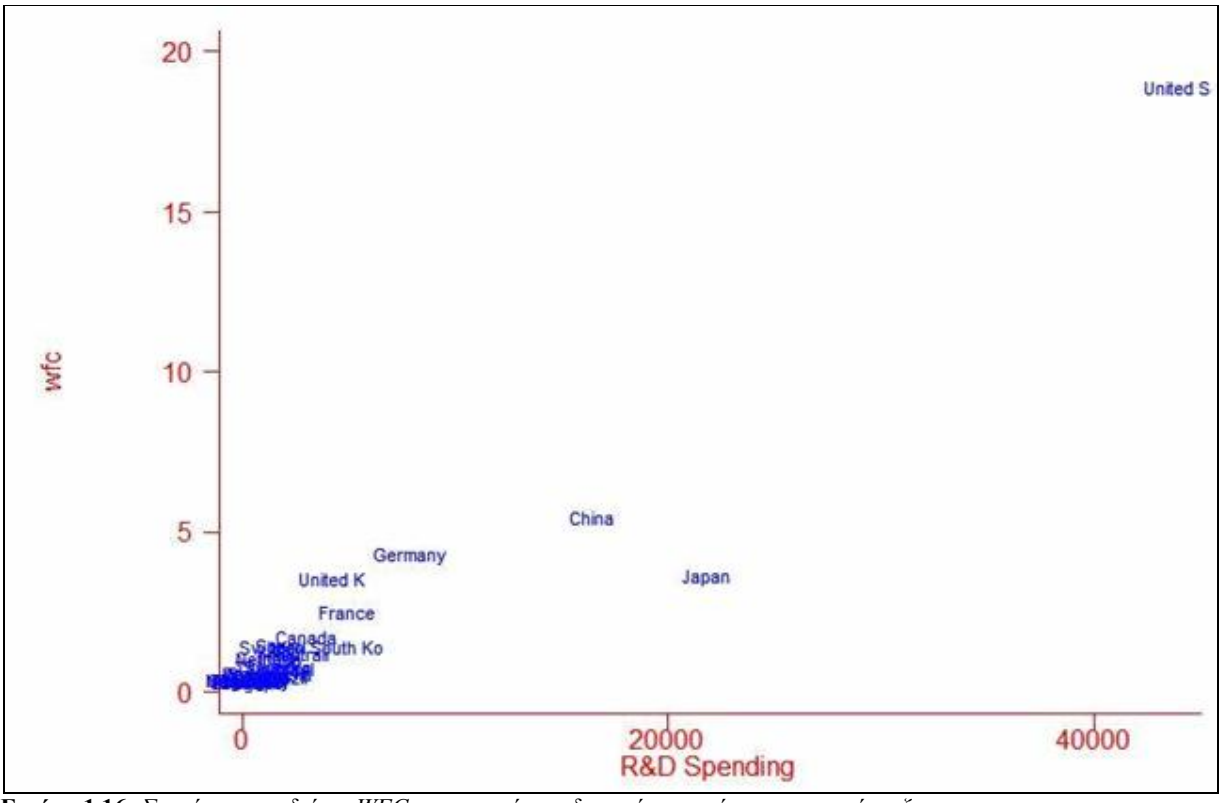
2013	COUNTRY	WFC	ARTICLE COUNT	2012 WFC	2012–2013 CHANGE IN WFC
1	United States	18,642.88	27,355	18,786.65	–0.8%
2	China	5,205.60	7,637	4,528.97	14.9%
3	Germany	4,076.97	8,669	4,038.30	1.0%
4	Japan	3,370.85	5,102	3,451.26	–2.3%
5	United Kingdom	3,290.35	7,373	3,259.46	0.9%
6	France	2,237.92	5,246	2,343.02	–4.5%
7	Canada	1,483.10	3,220	1,534.98	–3.4%
8	Spain	1,180.25	2,975	1,197.46	–1.4%
9	Switzerland	1,175.18	2,552	1,177.46	–0.2%
10	South Korea	1,150.52	1,953	1,193.02	–3.6%
11	Italy	1,075.12	3,089	1,084.27	–0.8%
12	Australia	943.94	2,448	864.60	9.2%
13	India	851.76	1,380	737.31	15.5%
14	Netherlands	763.24	2,221	765.03	–0.2%
15	Taiwan	543.18	937	595.99	–8.9%
16	Sweden	502.00	1,304	476.73	5.3%
17	Singapore	483.20	833	464.72	4.0%
18	Israel	472.35	1,008	520.05	–9.2%
19	Russia	344.26	1,058	298.26	15.4%
20	Belgium	327.25	1,019	347.22	–5.8%
21	Denmark	298.21	934	301.77	–1.2%
22	Austria	280.61	797	268.85	4.4%
23	Brazil	233.81	670	199.31	17.3%
24	Poland	216.35	689	176.78	22.4%
25	Finland	193.37	586	189.15	2.2%
26	Portugal	124.87	419	114.68	8.9%
27	Norway	123.62	371	142.00	–12.9%
28	New Zealand	118.93	307	127.84	–7.0%
29	Czech Republic	118.43	378	117.83	0.5%
30	Ireland	117.72	336	167.46	–29.7%
31	Argentina	105.66	304	93.08	13.5%
32	Greece	90.15	337	107.55	–16.2%
33	South Africa	81.81	441	58.02	41.0%

Εικόνα 1.14: *Η κατάταξη των χωρών με βάση το www.natureindex.com*

Βρίσκεται λοιπόν η Ελληνική επιστημονική έρευνα σε τόσο μεγάλη απαξίωση; Μια προσεκτικότερη ματιά στους πίνακες δείχνει κάτι απλό: η σειρά των χωρών στη λίστα θυμίζει πάρα πολύ τη σειρά των χωρών με βάση το συνολικό ΑΕΠ τους, έναν δείκτη που αντικατοπτρίζει την οικονομική δραστηριότητα αλλά και το μέγεθος της κάθε χώρας (η Ελλάδα στη σχετική λίστα βρίσκεται, παρά την ύφεση, στην 43η θέση παγκοσμίως, http://data.worldbank.org/, στοιχεία 2012-2013). Κατά συνέπεια, ο συντελεστής συσχέτισης του δείκτη WFC με το ΑΕΠ των χωρών της ΕΕ, αλλά και με το συνολικό ποσό επένδυσης στην έρευνα για κάθε χώρα, είναι περίπου 95%. Όσο περισσότερα λεφτά έχεις, τόσο περισσότερη έρευνα υψηλής ποιότητας παράγεις. Καμία έκπληξη. Αξίζει να σημειωθεί, ότι αν η ανάλυση επαναληφθεί στις πρώτες 33 χώρες στη λίστα του Nature (που περιλαμβάνει πάνω-κάτω τις περισσότερες χώρες του ΟΟΣΑ), τα αποτελέσματα είναι παρόμοια.



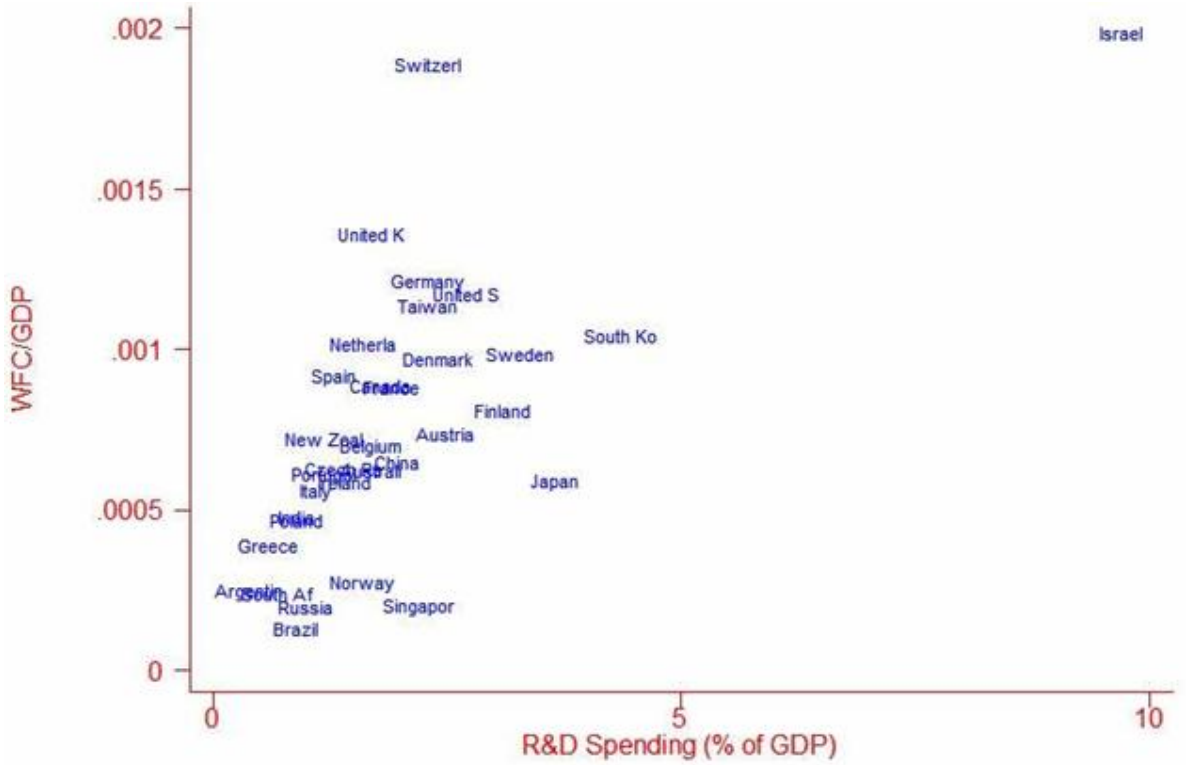
Εικόνα 1.15: *Συσχέτιση του δείκτη WFC με το ΑΕΠ της κάθε χώρας*



Εικόνα 1.16: *Συσχέτιση του δείκτη WFC με το ποσό που δαπανάται για έρευνα και ανάπτυξη*

Ένας σταθμισμένος δείκτης που είναι ανεξάρτητος από την ποσότητα της έρευνας και το μέγεθος της χώρας, βασίζεται στις αναφορές των δημοσιευμένων επιστημονικών εργασιών απο άλλες επιστημονικές εργασίες. Εκφράζεται ως ο λόγος των αναφορών προς τον αριθμό των εργασιών και κάνει μια εκτίμηση της απήχησης της έρευνας, χωρίς παραδοχές για την ποιότητα του περιοδικού δημοσίευσης. Πρόκειται δηλαδή για το ίδιο κριτήριο με αυτό που χρησιμοποίησε το ΕΚΤ στις αναλύσεις του, που αναφέραμε παραπάνω, αλλά για ευκολία χρησιμοποιήσαμε δεδομένα από άλλη πηγή (http://www.scimagojr.com, στοιχεία 2011-2013). Επειδή αυτός ο δείκτης είναι ήδη σταθμισμένος, δείχνει μια μάλλον αμελητέα συσχέτιση με το ΑΕΠ και το συνολικό ποσό επένδυσης στην έρευνα, αλλά μια αξιοπρόσεκτη συσχέτιση μόνο με το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα, ~70%. Όσο περισσότερα λεφτά βάζεις στην έρευνα αναλογικά με το ΑΕΠ, τόσο πιό μεγάλη απήχηση έχει η έρευνα που παράγεις.

Με βάση τις υψηλές συσχετίσεις αυτών των τριών ζευγαριών αξιολόγησης, μπορέσαμε να φτιάξουμε ένα απλό μοντέλο που προβλέπει το δείκτη WFC με βάση το ΑΕΠ και με βάση το συνολικό ποσό επένδυσης στην έρευνα και το δείκτη ετεροαναφορών ανά εργασία με βάση το ποσοστό του ΑΕΠ που επενδύει κάθε χώρα στην έρευνα. Από αυτό το μοντέλο μπορούμε να υπολογίσουμε την απόκλιση, θετική ή αρνητική, κάθε χώρας από το αναμενόμενο. Σύμφωνα με αυτή την ανάλυση, η Ελλάδα τα πάει κατά 29% καλύτερα απο το αναμενόμενο με βάση το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα (και βρίσκεται στην δεύτερη θέση) και κατά 48% καλύτερα απο το αναμενόμενο με βάση το συνολικό ποσό επένδυσης στην έρευνα, και βρίσκεται στην τρίτη θέση μαζί με το Ηνωμένο Βασίλειο (48%). Όταν όμως συγκρίνουμε την απόκλιση της Ελλάδας με βάση το συνολικό ΑΕΠ, αυτή βρίσκεται στο -15% από το αναμενόμενο, στη 19η θέση ανάμεσα στις χώρες της ΕΕ.



Εικόνα 1.17: *Συσχέτιση το λόγου WFC/ΑΕΠ με το ποσοστό του ΑΕΠ που δαπανάται στην έρευνα και ανάπτυξη*

Η ανάλυση αυτή είναι φυσικά περιορισμένη, και η επιλεγμένη μεθοδολογία καθώς και τα πρωτογενή δεδομένα μπορούν να αμφισβητηθούν. Επίσης, θα πρέπει να σημειώσουμε οτι σε όλες αυτές τις αναλύσεις η συσχέτιση (correlation) δεν σημαίνει υποχρεωτικά αιτιώδη συνάφεια (causation), αν και ειδικά στην περίπτωση του ζεύγους ΑΕΠ και ερευνητικής παραγωγή, υπάρχουν εμπειρικά και θεωρητικά δεδομένα που να υποστηρίζουν μια τέτοια σχέση. Σε κάθε περίπτωση, η ανάλυση αυτή είναι χρήσιμη και μπορεί να βοηθήσει στην εξαγωγή κάποιων συμπερασμάτων. Το μήνυμα από μια τέτοια ανάλυση είναι μάλλον απλό: καμία μα καμία βελτίωση δεν είναι δυνατόν να γίνει στην ποιότητα της Ελληνικής έρευνας εάν δεν αυξηθεί

το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα. Η Ελλάδα με βάση το απόλυτο ποσό αλλά και το ποσοστό του ΑΕΠ που επενδύεται στην έρευνα, τα πάει εξαιρετικά σε σχέση με τη συντριπτική πλειοψηφία άλλων Ευρωπαϊκών χωρών. Τα πάει όμως μάλλον άσχημα ή μέτρια σε σχεση με το (κουτσουρεμένο λόγω ύφεσης) ΑΕΠ της, για τον απλό λόγο οτι ελάχιστο ποσοστό του ΑΕΠ επενδύεται στην έρευνα. Απο τα μικρότερα στην Ευρώπη και λιγότερο απο το 1/3 του Ευρωπαϊκού στόχου για 2.1% επί του ΑΕΠ (0.69%) (Μπάγκος & Περράκης, 2014).

1.4.2. Η βιοπληροφορική έρευνα στην Ελλάδα

Ειδικότερα για την κοινότητα της βιοπληροφορικής στην Ελλάδα και την έρευνα που γίνεται στον τομέα αυτό, δεν υπάρχουν πολλά δημοσιευμένα δεδομένα. Τα τελευταία χρόνια όμως, έχουν γίνει σημαντικές προσπάθειες για την οργάνωση και την καταγραφή αυτής της δραστηριότητας στο πλαίσιο της ΕΕΥΒΒ. Η Ελληνική Εταιρεία Υπολογιστικής Βιολογίας και Βιοπληροφορικής (ΕΕΥΒΒ), http:///www.hscbb.gr είναι η επιστημονική εταιρεία στην οποία συμμετέχουν δεκάδες επιστήμονες από την Ελλάδα και την Κύπρο, οι οποίοι ασχολούνται με την Υπολογιστική Βιολογία και τη Βιοπληροφορική. Είναι η μοναδική επιστημονική εταιρεία (σωματείο) με το αντικείμενο αυτό στην Ελλάδα, λειτουργεί με τη νομική μορφή σωματείου από το 2009 και είναι συνδεδεμένο μέλος (affiliated) της αντίστοιχης διεθνούς ομοσπονδίας (International Society for Computational Biology, βλ. http://www.iscb.org/iscb-affiliates-europe#hellenic), ενώ συμμετέχει και σε διάφορες άλλες διεθνείς πρωτοβουλίες όπως το GOBLET (http://www.mygoblet.org) και το ELIXIR (https://www.elixir-europe.org/).

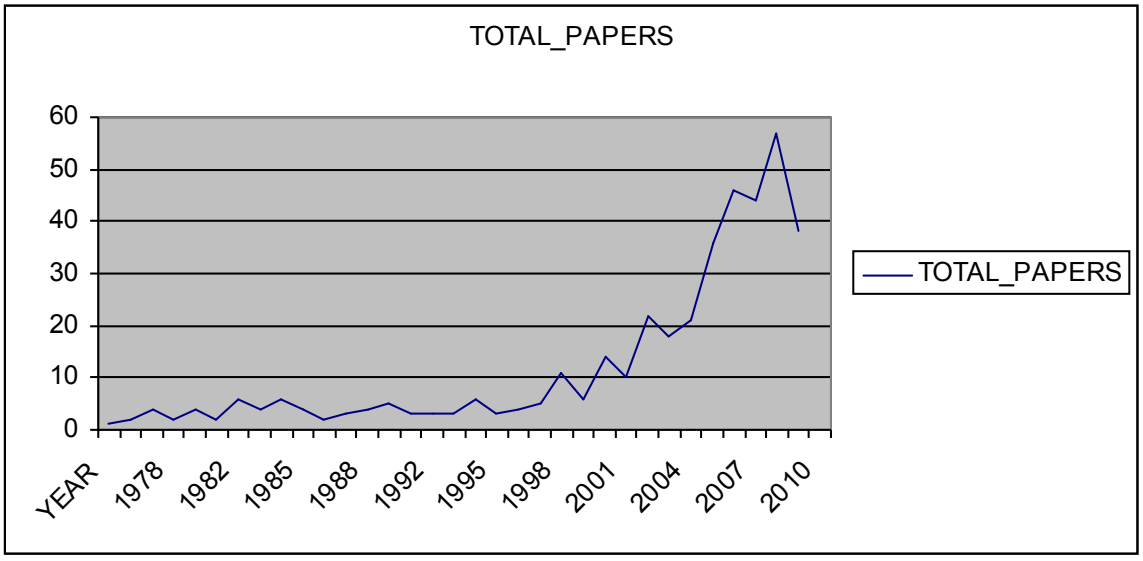
Τα συνέδρια της ΕΕΥΒΒ, διεξάγονται σε ετήσια βάση με μεγάλη επιτυχία και προβολή (π.χ. http://hscbb11.hscbb.gr, http://hscbb12.hscbb.gr κ.ο.κ.) και συμμετέχουν ως προσκεκλημένοι σε αυτά ομιλητές, που είναι 'μεγάλα ονόματα' από το εξωτερικό. Προηγούμενα συνέδρια έχουν γίνει σε πόλεις με σχετικά με το επιστημονικό πεδίο πανεπιστημιακά τμήματα και ερευνητικά ιδρύματα, π.χ. στην Αθήνα, στην Πάτρα, στο Ηράκλειο, στην Αλεξανδρούπολη και στη Λαμία. Αξίζει να σημειωθεί, ότι στα προηγούμενα συνέδρια, οι σύνεδροι ξεπερνούν τους 120 με συνεχώς αυξανόμενους αριθμούς, ενώ μεγάλο μέρος από αυτούς είναι προπτυχιακοί και μεταπτυχιακοί φοιτητές. Τα ενεργά μέλη της εταιρίας (μέλη ΔΕΠ, Ερευνητές, και γενικότερα κάτοχοι διδακτορικού) είναι περίπου 40 από όλη την Ελλάδα, αλλά ο συνολικός αριθμός είναι μεγαλύτερος καθώς κάποιοι συμμετέχουν μόνο περιστασιακά. Γενικά η ενεργή κοινότητα της βιοπληροφορικής στην Ελλάδα αριθμεί πάνω από 30 ερευνητικές ομάδες σε διάφορα πανεπιστήμια και ερευνητικά ιδρύματα, έστω και αν για πολλούς από αυτούς η βιοπληροφορική δεν είναι το μόνο ή το κύριο ερευνητικό αντικείμενο.

Σε μια προσπάθεια να καταγραφεί αναλυτικά η ερευνητική δραστηριότητα στην Ελλάδα, είχε γίνει μια σχετική εργασία που παρουσιάστηκε στο ετήσιο συνέδριο της ΕΕΥΒΒ το 2010 (Bagos, 2010). Σε αυτή την εργασία έγινε συστηματική προσπάθεια να αναλυθεί η βιβλιογραφία της βιοπληροφορικής συλλέγοντας όλες της εργασίες στις οποίες συμμετείχαν συγγραφείς με διεύθυνση εργασίας κάποιο Ελληνικό ίδρυμα και έγινε ανάλυση που αφορά τον αριθμό των εργασιών και των αναφορών τους, τους συγγραφείς, τα ινστιτούτα τους αλλά και τα ερευνητικά ενδιαφέροντα και τις τάσεις τους στην πορεία των χρόνων. Και σε αυτή την περίπτωση τα βασικά προβλήματα αυτών των εργασιών παραμένουν: δηλαδή το πώς θα συλλεχθεί ένα όσο το δυνατό μεγαλύτερο σύνολο από εργασίες από διάφορα περιοδικά (και κυρίως πώς θα διαχωριστούν με ακρίβεια τα περιοδικά βιοπληροφορικής), από ποια βάση δεδομένων θα γίνει η καταγραφή των δεδομένων, και με ποιον τρόπο θα γίνει η ανάλυση του κειμένου. Η επιλογή σε αυτή την περίπτωση ήταν να στηριχθούμε στην γενικότερη κατηγορία του ISI WoS, με τον τίτλο "MATHEMATICAL & COMPUTATIONAL BIOLOGY" και να συλλεχθούν όλα τα περιοδικά αυτής της κατηγορίας. Η κατηγορία αυτή περιέχει τα μεγαλύτερα αμιγώς βιοπληροφορικά περιοδικά, αλλά και κάποια αξιόλογα περιοδικα ιατρικής πληροφορικής και βιοστατιστικής. Σε μια προσπάθεια να δευρυνθεί το σύνολο δεδομένων, επιλέχθηκε και μια επιπλέον λίστα περιοδικών που έχουν ξεκάθαρη αναφορά στον τίτλο τους σε «βιοπληροφορική» ή «υπολογιστική βιολογία» αλλά και τα ειδικά τεύχη του Nucleic Acids Research που είναι αφιερωμένα σε εφαρμογές βιοπληροφορικής (web-server και database issues). Όλα τα περιοδικά αυτά, ήταν ενταγμένα στη βάση δεδομένων PUBMED (ακόμα και αν δεν ήταν στο WoS). Τα περιοδικά και των δύο κατηγοριών που χρησιμοποιήθηκαν στην ανάλυση δίνονται στον Πίνακα 1.4. Προφανώς, ένας σημαντικός περιορισμός του τρόπου αναζήτησης ήταν ότι πολλές εργασίες βιοπληροφορικής ή εργασίες που έκαναν εκταταμένη χρήση υπολογιστικών μεθόδων και είχαν δημοσιευτεί σε καθαρά βιολογικά περιοδικά (JMB, Plos Biology, Protein Engineering κ.ο.κ.) αλλά και εργασίες στα κορυφαία περιοδικά γενικού ενδιαφέροντος (Science, Nature) δεν

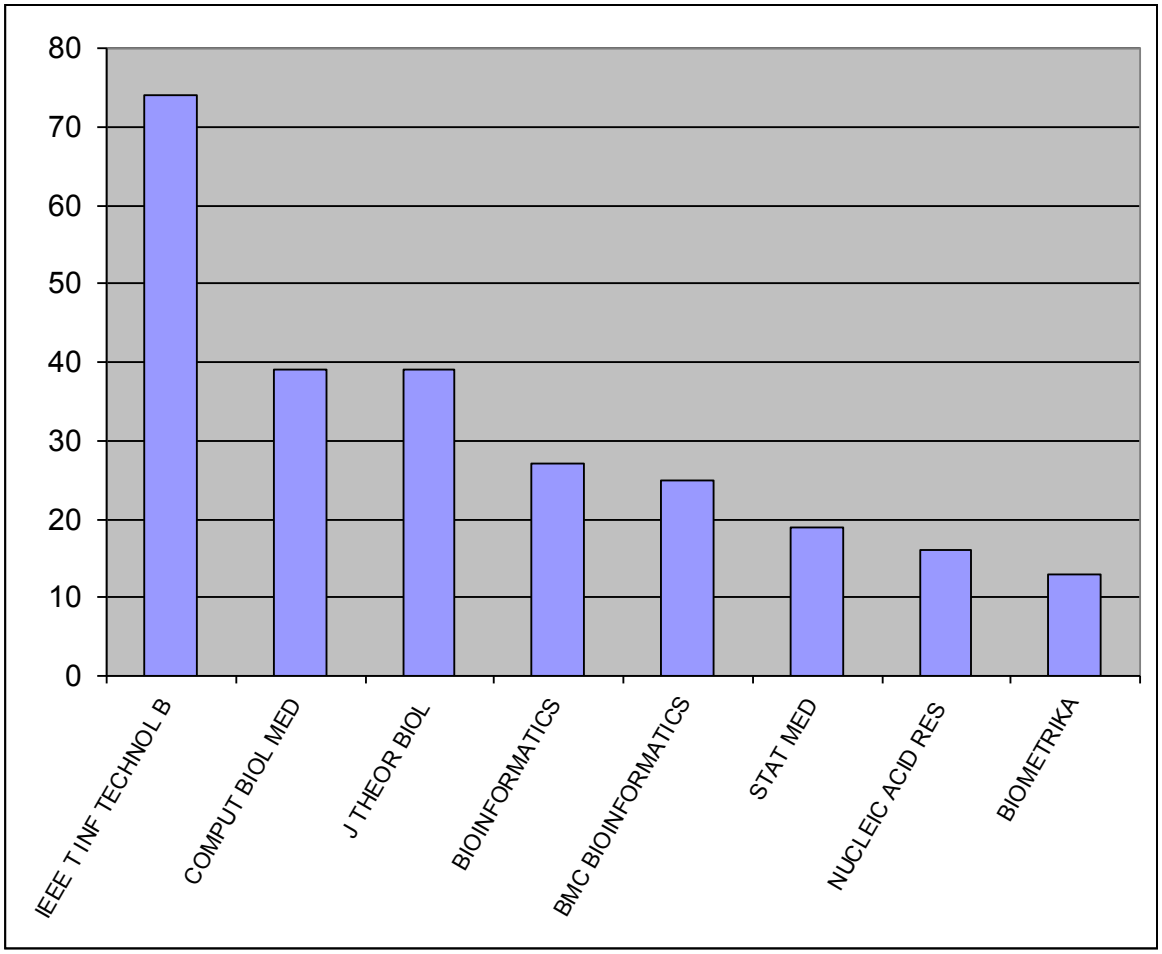
συμπεριλήφθηκαν στη μελέτη. Το ίδιο ισχύει και για αντίστοιχες εργασίες που έχουν δημοσιευθεί σε περιοδικά της πληροφορικής (Machine Learning, Pattern Recognition κ.ο.κ.). Κατά συνέπεια, τα πορίσματα αυτά αποτελούν υπο-εκτίμηση της πραγματικής διάστασης της σχετικής βιβλιογραφίας στην Ελλάδα. Επίσης, για μια σειρά σημαντικούς επιστήμονες του χώρου, μεγάλο μέρος από τις δημοσιευμένες εργασίες τους είχαν πραγματοποιηθεί όταν αυτοί εργάζονταν σε ιδρύματα του εξωτερικού, κατά συνέπεια δεν έχουν συμπεριληφθεί στην ανάλυση. Οι αναζητήσεις έγιναν με επιπλέον λέξη-κλειδί τη χώρα προέλευσης (GREECE ή CYPRUS), τα δεδομένα αποθηκεύτηκαν σε μια βάση δεδομένων SQL και αναλύθηκαν με στατιστικά εργαλεία και η ανάλυση του κειμένου έγινε με τη σχετική εφαρμογή του Yahoo, το γνωστό «Term Extraction web service» (http://developer.yahoo.com/search/content/V1/termExtraction.html), το οποίο απομονώνει από τις περιλήψεις τις σημαντικές λέξεις-κλειδιά (η επιλογή αυτή έγινε γιατί είδαμε ότι τα KEYWORDS του ίδιου του WoS είναι πολλές φορές αποπροσανατολιστικά ή πολύ γενικά).

PLOS Comput Biol	Journal of Computer-Aided Molecular Design
Bioinformatics	Nucleic Acids Research (web-server and database issues)
BMC Syst Biol	The Open Bioinformatics Journal
BMC Bioinformatics	Statistical Applications in Genetics and Molecular Biology
Biostatistics	Source Code for Biology and Medicine
J Theor Biol	Online Journal of Bioinformatics
Stat Method Med Res	Journal of Integrative Bioinformatics
IET Syst Biol	Journal of Bioinformatics and Computational Biology
J Comput Neurosci	International Journal of Data Mining and Bioinformatics
J Mol Graph Model	International Journal of Computational Biology and Drug Design
Stat Med	International Journal of Bioinformatics Research and Applications
Biometrika	In Silico Biology
Evol Bioinform	Genomics, Proteomics & Bioinformatics
B Math Biol	Genome Informatics
Biometrics	EURASIP Journal on Bioinformatics and Systems Biology
Algorithm Mol Biol	Current Bioinformatics
Med Biol Eng Comput	BioData Mining
J Math Biol	Advances and Applications in Bioinformatics and Chemistry
IEEE T Inf Technol B	Applied Bioinformatics
J Comput Biol	International Journal of Bioinformatics
Curr Bioinform	Bioinformation
SAR QSAR Environ Res	Pac Symp Biocomput
Math Biosci	Database
Comput Biol Med	Genome Res
Biometrical J	BMC Genomics
Math Med Biol	
Int J Data Min Bioin	
J Agr Biol Envir St	
J Biol Syst	

Πίνακας 1.4: *Τα περιοδικά που χρησιμοποιήθηκαν στη δική μας ανάλυση για τη βιοπληροφορική στην Ελλάδα*



Εικόνα 1.18: *Χρονική εξέλιξη των δημοσιεύσεων βιοπληροφορικής στην Ελλάδα*



Εικόνα 1.19: *Τα περιοδικά με τις περισσότερες εργασίες που εντοπίστηκαν στη μελέτη για τη βιοπληροφορική στην Ελλάδα*

Η ανάλυση έδωσε 405 εργασίες οι οποίες είχαν πραγματοποιηθεί από το 1976 μέχρι τις αρχές του 2010. Η αύξηση είναι, όπως και στην περίπτωση της διεθνούς βιβλιογραφίας, εκθετική μετά το 1999, καθώς

μέχρι εκείνη τη χρονιά είχαμε περίπου 5 εργασίες το χρόνο (Εικόνα 1.18). Συνολικά, στις 405 εργασίες είχαν συμμετάσχει 681 διαφορετικοί συγγραφείς (1,68 συγγραφείς ανά εργασία). Ο σχετικά μικρός αριθμός συγγραφέων ανά εργασία (σε σχέση με το αναμενόμενο), δικαιολογείται αν αναλογιστούμε ότι στη μελέτη περιλαμβάνονται και πολλές εργασίες, μαθηματικής βιολογίας και βιοστατιστικής, οι οποίες έχουν λίγους συγγραφείς, πολλές φορές και μόνο έναν. Από τους 681 συγγραφείς, οι 636 είχαν εμπλακεί σε 3 το πολύ εργασίες ενώ οι 45 είχαν συμμετάσχει σε περισσότερες, ενώ μόλις 18 είχαν συμμετάσχει σε περισσότερες από 9 εργασίες. Τα δεδομένα αυτά είναι συμβατά με τα αντίστοιχα στη διεθνή βιβλιοφραφία που αναλύσαμε παραπάνω (ειδικά αν αναλογιστούμε ότι κάποιοι συγγραφείς που εμφανίζονται με λίγες εργασίες είχαν δημοσιεύσει περισσότερες όταν εργάζονταν στο εξωτερικό). Συνολικά, εντοπίστηκαν 63 διαφορετικά εκπαιδευτικά και ερευνητικά ιδρύματα, τα περισσότερα εκ των οποίων ήταν πανεπιστήμια.

Από τους συγγραφείς, ο πιο επιδραστικός είναι ο καθηγητής Σταύρος Χαμόδρακας από το ΕΚΠΑ, τόσο σε απόλυτο αριθμό εργασιών, όσο και σε αριθμό αναφορών αλλά και συνολικό Impact Factor. Το γεγονός αυτό είναι κάτι αναμενόμενο, καθώς ο καθηγητής Χαμόδρακας ήταν από τους πρώτους που ασχολήθηκαν με τη βιοπληροφορική στην Ελλάδα και ήταν ο συγγραφέας των 2 από τις 3 εργασίες που είχαν δημοσιευθεί από ερευνητές Ελληνικών ιδρυμάτων στο περιοδικό Computer Applications in Biosciences, τον πρόδρομο του πιο γνωστού περιοδικού στο χώρο, του Bioinformatics. Οι εργασίες αυτές, είχαν τίτλο «A protein secondary structure prediction scheme for the IBM PC and compatibles» του 1988 (3208182) και «PBM: a software package to create, display and manipulate interactively models of small molecules and proteins on IBM-compatible PCs.» του 1995 (με Perrakis A, Constantinides C, Athanasiades A) 7620985, εργασίες που δικαίως μπορούν να χαρακτηριστούν τόσο ως οι πρώτες καθαρά βιοπληροφορικές εργασίες στην Ελλάδα, όσο και αντιπροσωπευτικές του ερευνητικού κλίματος της εποχής εκείνης.

Η πιο επιδραστική εργασία πριν το 2000 ήταν η εργασία των Fickett JW, Hatzigeorgiou AC. με τίτλο «Eukaryotic promoter recognition» στο Genome Res. 2^{η} ήταν η εργασία των Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA με τίτλο «CAST: an iterative algorithm for the complexity analysis of sequence tracts» στο Bioinformatics και 3^{η} η εργασία των Pavlou S, και Kevrekidis IG με τίτλο «Microbial predation in a periodically operated chemostat- a global study of the interaction between natural and externally imposed frequencies», η οποία δημοσιεύθηκε στο Math Biosci. Την περίδο 2001-2005, πιο επιδραστική εργασία ήταν των Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D et al. (με συμμετοχή του Έλληνα ερευνητή V Aidinis) με τίτλο «Targeting a complex transcriptome: The construction of the mouse full-length cDNA encyclopedia» στο περιοδικό Genome Res, 2^{η} ήταν η εργασία των Patrinos GP, Giardine B, Riemer C, Miller W, Chui DHK, Anagnou NP, Wajcman H, Hardison RC με τίτλο «Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies» στο Nucleic Acids Res και 3^{η} η εργασία των Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ με τίτλο «PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins», επίσης στο Nucleic Acids Res. Τέλος, την περίοδο 2006-2010, η πιο επιδραστική εργασία ήταν των Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC με τίτλο «The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide» στο Nucleic Acids Res, στη 2^{η} θέση ήταν η εργασία της ίδια ομάδας (Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC) με τίτλο «The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata» επίσης στο Nucleic Acids Res, ενώ στην 3^{η} θέση ήταν η εργασία των Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z et al. με τίτλο «BioMagResBank» και αυτή στο ίδιο περιοδικό (Nucleic Acids Res).

Η ανάλυση των 20 εργασιών με τις περισσότερες αναφορές, μας δείχνει ότι ανάμεσά τους περιλαμβάνονται 7 εργασίες που περιγράφουν βιολογικές βάσεις δεδομένων και 4 εργασίες με web-servers ή μεθόδους πρόγνωσης. Όμοια, ανάλυση των 20 εργασιών με τις περισσότερες αναφορές/χρόνο δείχνει ότι ανάμεσά τους βρίσκονται 8 εργασίες που περιγράφουν βιολογικές βάσεις δεδομένων και 9 εργασίες με web-servers ή μεθόδους πρόγνωσης. Μια στατιστική ανάλυση έδειξε ότι οι σημαντικότεροι παράγοντες που «προβλέπουν» τον αριθμό αναφορών που θα πάρει μια εργασία (εκτός από το Impact Factor του περιοδικού, κάτι το οποίο είναι αναμενόμενο), είναι ο αριθμός των συγγραφέων (όσο περισσότεροι, τόσο το καλύτερο), η συμμετοχή συγγραφέων από το εξωτερικό και το αν η εργασία περιγράφει μια βιολογική βάση δεδομένων ή όχι. Ο αριθμός των συγγραφέων και η συμμετοχή ξένων επιστημόνων φαίνεται ότι είναι παράγοντες ενδεικτικοί της ποιότητας της εργασίας, αν και υπάρχουν υπόνοιες για κάποιου είδους bias (π.χ. οι εργασίες με ξένους επιστήμονες ενδέχεται να θεωρούνται καλύτερες και γι' αυτό να αναφέρονται περισσότερο). Η

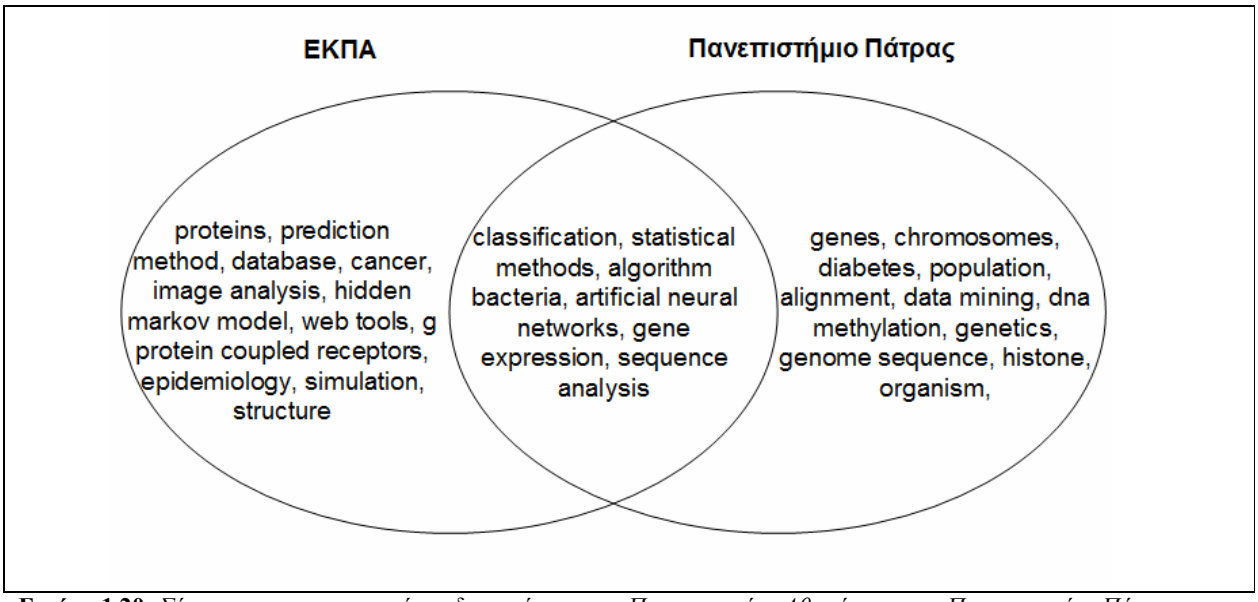
τόσο μεγάλη επιρροή που φαίνεται ότι έχουν ο βιολογικές βάσεις δεδομένων, ειδικά μετά το 2005, είναι σε συμφωνία με τα όσα είδαμε στις προηγούμενες ενότητες σχετικά με τη διεθνή βιβλιογραφία.

Σε σχέση με τα ινστιτούτα από τα οποία προήλθαν οι εργασίες βιοπληροφορικής, στην πρώτη θέση τόσο όσον αφορά τον απόλυτο αριθμό των εργασιών, αλλά και των αναφορών και του Impact Factor, βρίσκεται το ΕΚΠΑ και ακολουθούν το Πανεπιστήμιο Πάτρας, το Πανεπιστήμιο Ιωαννίνων, το ΑΠΘ και το ΕΜΠ (Πίνακας 1.5). Τη δεκάδα συμπληρώνουν το Πανεπιστήμιο Κρήτης και το Πανεπιστήμιο Κύπρου, ενώ οι μόνες παρουσίες ερευνητικών κέντρων στην πρώτη δεκάδα είναι το ΙΤΕ (6η θέση), το ΕΚΕΦΕ Δημόκριτος (8η θέση) και το ΙΙΒΕΑΑ (9η θέση). Αυτή η κατάταξη αφορά διαχρονικά την εποχή στην οποία δημοσιεύτηκε η εργασία. Όταν η ίδια κατάταξη γίνει με βάση την τωρινή θέση που κατέχουν οι επιστήμονες (με βάση τις τωρινές θέσεις εργασίας των 18 επιστημόνων με τις περισσότερες εργασίες), η σειρά αλλάζει λίγο και είναι αυτή που φαίνεται στον Πίνακα. Παρατηρούμε, ότι η γενική εικόνα δεν έχει αλλάξει πολύ, για παράδειγμα το ΕΚΠΑ εξακολουθεί να είναι πρώτο, τα Πανεπιστήμια Ιωαννίνων και Πάτρας, αλλά και το ΕΜΠ είναι μέσα στην πρώτη πεντάδα κ.ο.κ. Παρόλα αυτά, βλέπουμε τη δυναμική εμφάνιση του Πανεπιστημίου Στερεάς Ελλάδας (3η θέση), αλλά και την είσοδο δύο νέων ιδρυμάτων στην πρώτη δεκάδα, του ΑΛ. ΦΛΕΜΙΝΓΚ (6η θέση) αλλά και του ΤΕΙ Αθήνας (9η θέση).

Ίδρυμα	**Εργασίες**	**Αναφορές**	**Impact Factor**
University of Athens	53	706	159,86
University of Ioannina	41	192	65,53
University of Central Greece	34	498	97,39
Natl Tech Univ Athens	28	139	54,21
University of Patras	21	66	48,17
BSRC Alexander Fleming	18	362	103,11
Aristotle Univ Thessaloniki	16	49	25,66
Natl Ctr Sci Res Demokritos	15	110	35,9
Tech Educ Inst Athens	14	46	28,73
Acad Athens|Biomed Res Fdn	12	13	35,33
Ctr Res & Technol Hellas CERTH	9	116	32,87
University of Thessaly	9	70	17,46
University of Cyprus	8	139	32,53
Tech Educ Inst Lamia	6	42	11,47
Democritus Univ Thrace	6	31	15,44

Πίνακας 1.5: *Τα κυριότερα ιδρύματα που εντοπίστηκαν στη μελέτη μας, με τις εργασίες, τις αναφορές και το συνολικό δείκτη επιρροής.*

Συμπερασματικά, και παρ' όλους τους περιορισμούς της μελέτης αυτής τους οποίους αναλύσαμε παραπάνω, μπορούμε να πούμε ότι η δραστηριότητα στον Ελληνικό χώρο την τελευταία 15ετία είναι ιδιαίτερα αυξημένη και έχει αρχίσει να σχηματίζεται η κρίσιμη μάζα επιστημόνων οι οποίοι θα προωθήσουν το πεδίο. Η διεπιστημονικότητα στην προέλευση αυτών των επιστημόνων είναι εμφανής, τόσο όταν αναλογιστούμε το υπόβαθρο των παλαιότερων ερευνητών στο χώρο, όσο και βλέποντας τα ιδρύματα στα οποία διεξάγεται η έρευνα αυτή. Τα παλαιότερα και μεγαλύτερα πανεπιστήμια, όπως ήταν αναμενόμενο, κυριαρχούν στο χώρο, αλλά τα νεότερα περιφερειακά πανεπιστήμια έχουν έντονη παρουσία τα τελευταία χρόνια.



Εικόνα 1.20: *Σύγκριση των ερευνητικών ενδιαφερόντων του Πανεπιστημίου Αθηνών και του Πανεπιστημίου Πάτρας.*

1.4.3. Η εκπαίδευση στη βιοπληροφορική στην Ελλάδα

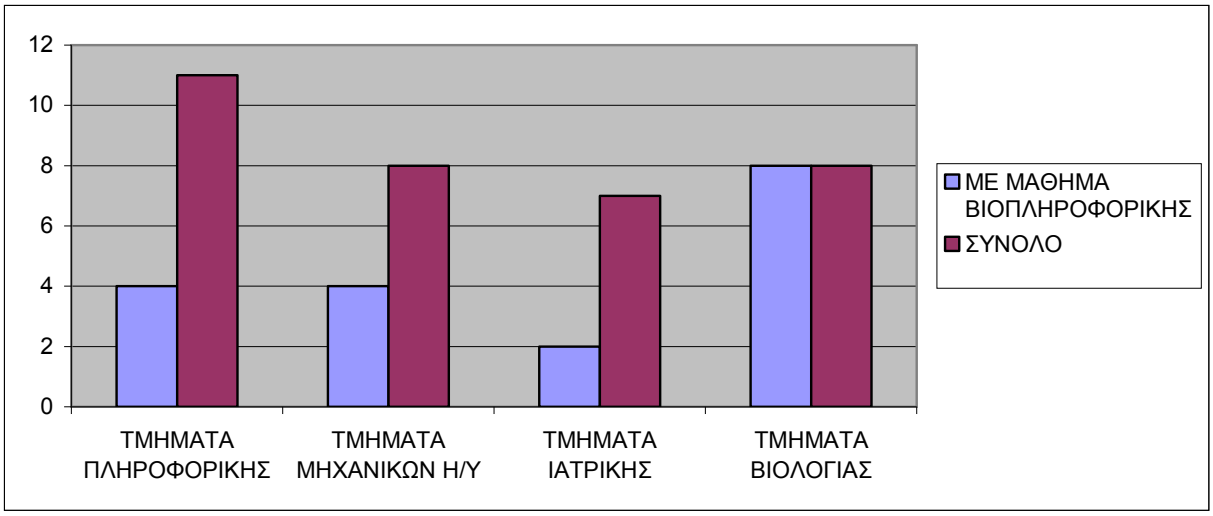
Τέλος, έχει ιδιαίτερη αξία να δούμε εκτός από την ερευνητική και την εκπαιδευτική δραστηριότητα στον Ελληνικό χώρο, τόσο σε προπτυχιακό όσο και σε μεταπτυχιακό επίπεδο (τουλάχιστον στο πλαίσιο κάποιου οργανωμένου προγράμματος σπουδών). Εκτεταμένη αναζήτηση στις ιστοσελίδες και στα προγράμματα σπουδών των Ελληνικών Τμημάτων Βιολογίας (και των υπόλοιπων συναφών τμημάτων βιολογικών επιστημών), Ιατρικής, Πληροφορικής, αλλά και Μηχανικών Η/Υ, δειχνει ότι Βιοπληροφορική διδάσκεται σε προπτυχιακό επίπεδο σε 18 τμήματα σε όλη την Ελλάδα (Πίνακας 1.6). Οκτώ από αυτά είναι βιολογικά ή συναφή τμήματα, τέσσερα είναι τμήματα Μηχανικών Η/Υ, τέσσερα είναι τμήματα Πληροφορικής και δύο είναι Ιατρικές σχολές. Βλέπουμε λοιπόν ότι σε όλα σχεδόν τα βιολογικής κατεύθυνσης τμήματα της χώρας, υπάρχει σχετικό μάθημα βιοπληροφορικής στο πρόγραμμα σπουδών. Αντίθετα, το ίδιο συμβαίνει στα λιγότερο από τα μισά τμήματα Πληροφορικής, Μηχανικών Η/Υ αλλά και Ιατρικής. Στα 11 από τα 18 τμήματα, υπηρετεί μέλος ΔΕΠ με γνωστικό αντικείμενο Βιοπληροφορική ή συναφές (π.χ. Υπολογιστική Βιολογία), σε 3 από τα 18 τμήματα υπηρετεί μέλος ΔΕΠ το οποίο έχει τη βιοπληροφορική στα κύρια ερευνητικά του ενδιαφέροντα ενώ στα υπόλοιπα 3 τμήματα, δεν υπάρχει σχετικό μέλος ΔΕΠ (στο Τμήμα Βιολογίας του ΕΚΠΑ ο καθ. Σ. Χαμόδρακας αφυπηρέτησε, ενώ στο Τμήμα Πληροφορικής του Πανεπιστημίου Πειραιά το μάθημα δεν προσφέρεται καν στους φοιτητές).

Ενδιαφέρουσα περίπτωση είναι το Πανεπιστήμιο Θεσσαλίας, όπου τέσσερα διαφορετικά τμήματα, διαφορετικών σχολών, έχουν την βιοπληροφορική στο πρόγραμμα σπουδών τους, ενώ και στα τέσσερα υπηρετεί μέλος ΔΕΠ με αντίστοιχο γνωστικό αντικείμενο. Στο Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική, διδάσκονται μάλιστα 3 μαθήματα βιοπληροφορικής ενώ υπάρχουν στο πρόγραμμα σπουδών και άλλα συναφή μαθήματα (Βιοστατιστική, Βιολογία, Βιοχημεία, Γενετική, αλλά και αρκετά μαθήματα Ιατρικής Πληροφορικής). Επίσης, το Τμήμα Μοριακής Βιολογίας του Δημοκρίτειου Πανεπιστημίου Θράκης είναι μια ειδική περίπτωση καθώς εκεί διδάσκονται 4 εξαμηνιαία μαθήματα βιοπληροφορικής και υπολογιστικής βιολογίας, τα περισσότερα από κάθε άλλο τμήμα. Ιδιαίτερη έμφαση στη Βιοπληροφορική δίνεται και στο Πανεπιστήμιο Κρήτης, όπου το μάθημα διδάσκεται στα Τμήματα Βιολογίας, Ιατρικής, και Επιστήμης Υπολογιστών, ενώ σε όλες τις περιπτώσεις στα τμήματα αυτά υπηρετεί μέλος ΔΕΠ με συναφές γνωστικό αντικείμενο. Ειδικά στο Τμήμα Επιστήμης Υπολογιστών διδάσκονται δύο σχετικά μαθήματα ενώ στους φοιτητές δίνεται η δυνατότητα να επιλέξουν και μαθήματα βιολογίας από το αντίστοιχο τμήμα. Στο Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών διδάσκεται βιοπληροφορική σαν επιλεγόμενο μάθημα τόσο στο τμήμα Βιολογίας όσο και στο Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ενώ στο Πανεπιστήμιο Πατρών, στο Τμήμα Βιολογίας και στο Τμήμα Μηχανικών Η/Υ και Πληροφορικής (δεν υπάρχει τμήμα Πληροφορικής). Τέλος, εντύπωση προκαλεί αρχικά η απουσία του ΕΜΠ και των Τμημάτων Πληροφορικής και Μηχανικών του ΑΠΘ από τη σχετική λίστα αλλά τα ευρήματα αυτά γίνονται κατανοητά αν

αναλογιστούμε τη χαμηλή συνεισφορά των ιδρυμάτων αυτών στην έρευνα στη Βιοπληροφορική, όπως είδαμε στην προηγούμενη παράγραφο. Στα ιδρύματα αυτά οι περισσότεροι ερευνητές που εμπλέκονται σε θέματα βιοϊατρικής πληροφορικής, ασχολούνται κυρίως με την Ιατρική Πληροφορική και την Πληροφορική Υγείας, τομείς που εφάπτονται μεν, αλλά δεν ταυτίζονται με τη Βιοπληροφορική.

Πανεπιστήμιο	**Τμήμα**	**Μαθήματα**
Πανεπιστήμιο Θεσσαλίας	Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών (*)	Βιοπληροφορική
	Τμήμα ιατρικής (*)	Βιοπληροφορική-Βιομετρία
	Τμήμα Βιοχημείας και Βιοτεχνολογίας (*)	Βιοπληροφορική
	Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική (*)	Βιοπληροφορική Ι, Βιοπληροφορική ΙΙ, Ειδικά Θέματα Βιοπληροφορικής και Βιοηθική
Πανεπιστήμιο Κρήτης	Τμήμα Επιστήμης Υπολογιστών (*)	Αλγόριθμοι στη βιοπληροφορική, Εισαγωγή στον προγραμματισμό για Βιοπληροφορική
	Τμήμα Ιατρικής (*)	Εισαγωγή στη Βιοπληροφορική
	Τμήμα Βιολογίας (*)	Υπολογιστική Βιολογία
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών	Τμήμα Βιολογίας	Βιοπληροφορική
	Τμήμα Πληροφορικής και Τηλεπικοινωνιών (**)	Αλγόριθμοι Βιοπληροφορικής
Πανεπιστήμιο Πατρών	Τμήμα Βιολογίας (**)	Βιοπληροφορική
	Τμήμα μηχανικών Η/Υ και Πληροφορικής (**)	Εισαγωγή στη Βιοπληροφορική
Δημοκρίτειο Πανεπιστήμιο Θράκης	Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών	Βιοπληροφορική
	Τμήμα Μοριακής Βιολογίας και Γενετικής (*)	Εισαγωγή στην Υπολογιστική Βιολογία, Βιοπληροφορική, Ειδικά θέματα Βιοπληροφορικής, Ειδικά Θέματα Υπολογιστικής Βιολογίας
Πανεπιστήμιο Ιωαννίνων	Τμήμα Βιολογικών εφαρμογών και Τεχνολογιών (*)	Βιοπληροφορική, Ειδικά θέματα Βιοπληροφορικής
Πανεπιστήμιο Δυτικής Μακεδονίας	Τμήμα Μηχανικών Πληροφορικής και Τηλεπικοινωνιών (*)	Βιοπληροφορική
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης	Τμήμα Βιολογίας (*)	Βιοπληροφορική
Γεωπονικό Πανεπιστήμιο Αθηνών	Τμήμα Βιοτεχνολογίας (*)	Βιοπληροφορική
Πανεπιστήμιο Πειραιά	Τμήμα Πληροφορικής	Βιοπληροφορική

Πίνακας 1.6: *Τα πανεπιστημιακά τμήματα στα οποία διδάσκονται μαθήματα βιοπληροφορικής. Με (*) συμβολίζονται τα τμήματα στα οποία υπηρετεί μέλος ΔΕΠ με συναφές γνωστικό αντικείμενο.*



Εικόνα 1.21: *Τα τμήματα που έχουν μάθημα βιοπληροφορικής σε σχέση με τα υπόλοιπα συναφή τμήματα σε Ελληνικά Πανεπιστήμια.*

Όσον αφορά τη μεταπτυχιακή εκπαίδευση, αυτή τη στιγμή υπάρχουν στη χώρα 6 προγράμματα μεταπτυχιακών σπουδών (ΠΜΣ) που οδηγούν σε μεταπτυχιακό δίπλωμα ειδίκευσης (ΜΔΕ) στης βιοπληροφορική ή σε συναφείς ειδικότητες (δεν αναφέρονται τα προγράμματα βιοστατιστικής και ιατρικής πληροφορικής). Τα προγράμματα αυτά δίνονται στον Πίνακα 1.7.

Πανεπιστήμιο	**Τμήμα**	**ΜΔΕ**
Πανεπιστήμιο Θεσσαλίας	Τμήμα Πληροφορικής με εφαρμογές στη Βιοϊατρική και Τμήμα Πληροφορικής	Πληροφορική και Υπολογιστική Βιοϊατρική (με ροή «Υπολογιστική Ιατρική και Βιολογία»)
	Τμήμα ιατρικής	Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών	Τμήμα Βιολογίας	Βιοπληροφορική
	Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ΤΕΙ Αθήνας	Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία (με ροή «Βιοπληροφορική»)
Πανεπιστήμιο Πατρών	Τμήματα Ιατρικής, Βιολογίας, Φυσικής, Φαρμακευτικής και Μηχανικών Η/Υ και Πληροφορικής	Πληροφορική Επιστημών Ζωής (με ροή «Βιοπληροφορική»)
Γεωπονικό Πανεπιστήμιο Αθηνών	Βιοτεχνολογίας	Βιολογία Συστημάτων

Πίνακας 1.7: *Τα μεταπτυχιακά συναφή με τη βιοπληροφορική στα Ελληνικά Πανεπιστήμια.*

Από τα ΠΜΣ αυτά, τα παλιότερα είναι το ΠΜΣ «Βιοπληροφορικής» του Τμήματος Βιολογίας του ΕΚΠΑ, και το διατμηματικό ΠΜΣ «Πληροφορική Επιστημών Ζωής» (με ροή «Βιοπληροφορική») το οποίο συνδιοργανώνεται από τα Ιατρικής, Βιολογίας, Φυσικής, Φαρμακευτικής και Μηχανικών Η/Υ και Πληροφορικής του Πανεπιστημίου Πατρών. Τα δύο αυτά μεταπτυχιακά ιδρύθηκαν το 2003, ενώ λίγα χρόνια αργότερα ιδρύθηκε το ΠΜΣ «Τεχνολογίες Πληροφορικής στην Ιατρική και τη Βιολογία» (με ροή «Βιοπληροφορική») από το Τμήμα Πληροφορικής και Τηλεπικοινωνιών του ΕΚΠΑ σε συνεργασία με το ΤΕΙ Αθήνας. Βλέπουμε, ότι η έντονη ερευνητική δραστηριότητα των δύο αυτών ιδρυμάτων οδήγησε και στη δημιουργία των πρώτων ΠΜΣ στην Ελλάδα. Παρόλα αυτά, στην περίπτωση της Πάτρας είδαμε μια διατμηματική συνεργασία, ενώ στην περίπτωση του ΕΚΠΑ κάθε τμήμα οργάνωσε το δικό του πρόγραμμα σπουδών. Ιδιαίτερη περίπτωση είναι και πάλι το Πανεπιστήμιο Θεσσαλίας, στο οποίο λειτουργούν εδώ και λίγο καιρό δύο διαφορετικά μεταπτυχιακά προγράμματα σπουδών, το «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική» από το Τμήμα Ιατρικής στη Λάρισα, και το «Πληροφορική και Υπολογιστική Βιοϊατρική» (με ροή «Υπολογιστική Ιατρική και Βιολογία») από τα Τμήματα Πληροφορικής

με εφαρμογές στη Βιοϊατρική και Πληροφορικής, στη Λαμία. Εδώ, η γεωγραφική απομόνωση και το γεγονός ότι τα τμήματα βρίσκονται σε διαφορετικές πόλεις οδήγησε στη δημιουργία διαφορετικών προγραμμάτων σπουδών. Παρόλα αυτά, είναι εμφανές και εδώ ότι η έντονη ερευνητική παρουσία του Πανεπιστημίου Θεσσαλίας αλλά και η ύπαρξη μελών ΔΕΠ με συναφές με τη Βιοπληροφορική γνωστικό αντικείμενο έχει παίξει καταλυτικό ρόλο. Τελευταία προσθήκη είναι το ΠΜΣ «Βιολογία Συστημάτων», στο Τμήμα Βιοτεχνολογίας του Γεωπονικού Πανεπιστημίου Αθηνών, μεταπτυχιακό πρόγραμμα που εντάσσεται στα ερευνητικά ενδιαφέροντα του τμήματος και είναι συμβατό με την προπτυχιακή εκπαίδευση στο ίδρυμα αυτό. Εντύπωση προκαλεί η απουσία ΠΜΣ από το Πανεπιστήμιο Ιωαννίνων αλλά ακόμα περισσότερο από το Πανεπιστήμιο Κρήτης, τα οποία όπως είδαμε διαθέτουν και προσωπικό σε συναφή γνωστικά αντικείμενα αλλά και έχουν να επιδείξουν ερευνητική δραστηριότητα στον τομέα. Πιθανότατα, οι ανάγκες μεταπτυχιακής εκπαίδευσης στα ιδρύματα αυτά καλύπτονται από άλλα συναφή ή πιο γενικά προγράμματα σπουδών και οι φοιτητές που επιθυμούν να ασχοληθούν με βιοπληροφορική, βρίσκουν διέξοδο σε επίπεδο εκπόνησης διδακτορικής διατριβής.

Βιβλιογραφία

Altman, R. B. (1998). A curriculum for bioinformatics: the time is ripe. *Bioinformatics, 14*(7), 549-550.

Bagos, P. G. (2010). *Bioinformatics and Computational Biology in Greece: a bibliometric study*. Paper presented at the 5th Conference of HSCBB (HSCBB10), Alexandroupolis.

Chalmers, A. (1999). *What Is This Thing Called Science?* (3rd revised edition ed.). Hackett: University of Queensland Press, Open University press.

Ditty, J. L., Kvaal, C. A., Goodner, B., Freyermuth, S. K., Bailey, C., Britton, R. A., . . . Kerfeld, C. A. (2010). Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol, 8*(8), e1000448.

Eddy, S. R. (2005). "Antedisciplinary" science. *PLoS Comput Biol, 1*(1), e6.

Floriano, W. B. (2008). A portable bioinformatics course for upper-division undergraduate curriculum in sciences. *Biochem Mol Biol Educ, 36*(5), 325-335.

Hagen, J. B. (2000). The origins of bioinformatics. *Nat Rev Genet, 1*(3), 231-236.

Honts, J. E. (2003). Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. *Cell Biol Educ, 2*(4), 233-247.

King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6997), 311-316.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med, 40*(4), 346-358.

Molenberghs, G. (2005). Biometry, biometrics, biostatistics, bioinformatics,..., bio-X. *Biometrics, 61*(1), 1-9.

Ouzounis, C. (2000). Two or three myths about bioinformatics. *Bioinformatics, 16*(3), 187-189.

Ouzounis, C. (2002). Bioinformatics and the theoretical foundations of molecular biology. *Bioinformatics, 18*(3), 377-378.

Ouzounis, C. A. (2012). Rise and demise of bioinformatics? Promise and progress. *PLoS Comput Biol, 8*(4), e1002487.

Ouzounis, C. A., & Valencia, A. (2003). Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics, 19*(17), 2176-2190.

Patra, S. K., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics, 67*(3), 477-489.

Perez-Iratxeta, C., Andrade-Navarro, M. A., & Wren, J. D. (2007). Evolving research trends in bioinformatics. *Brief Bioinform, 8*(2), 88-95.

Rebholz-Schuhman, D., Cameron, G., Clark, D., van Mulligen, E., Coatrieux, J. L., Del Hoyo Barbolla, E., . . . Van der Lei, J. (2007). SYMBIOmatics: synergies in Medical Informatics and Bioinformatics--exploring current scientific literature for emerging topics. *BMC Bioinformatics, 8 Suppl 1*, S18.

Roberts, R. J. (2000). The early days of bioinformatics publishing. *Bioinformatics, 16*(1), 2-4.

Searls, D. B. (2010). The roots of bioinformatics. *PLoS Comput Biol, 6*(6), e1000809.

Searls, D. B. (2012). An online bioinformatics curriculum. *PLoS Comput Biol, 8*(9), e1002632.

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the Association for Information Science and Technology, 65*(2), 352-371.

Trifonov, E. N. (2000). Earliest pages of bioinformatics. *Bioinformatics, 16*(1), 5-9.

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., & Schneider, M. V. (2014). Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput Biol, 10*(3), e1003496.

Wingender, E. (1998). ISB: Just Another Journal? *In Silico Biol, 1*(1), 1-4.

Yan, B., Ban, K. H., & Tan, T. W. (2014). Integrating translational bioinformatics into the medical curriculum. *Int J Med Educ, 5*, 132-134.

Μαυρικάκη, Ε., Γκούβρα, Μ., & Καμπούρη, Α. (2014). *Βιολογία Γ Γυμνασίου*. Αθήνα: ΟΕΔΒ

Μπάγκος, Π., & Περράκης, Α. (2014, 25/11/2014). Το Ελληνικό παράδοξο στην επιστημονική έρευνα. *Το ΒΗΜΑ*.

Σαχίνη, Ε., Μάλλιου, Ν., & Χούσος, Ν. (2012). Ελληνικές Επιστημονικές Δημοσιεύσεις 1996-2010: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά Retrieved from http://reports.metrics.ekt.gr/

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραϊσκος, Δ. (2012). Ελληνικές Επιστημονικές Δημοσιεύσεις 2000-2010 - Τομέας Βιοεπιστημών Retrieved from http://metrics.ekt.gr/el/node/15

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραϊσκος, Δ. (2013). Ελληνικές Επιστημονικές Δημοσιεύσεις 1996-2010: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά - Scopus Retrieved from http://report03.metrics.ekt.gr

Σαχίνη, Ε., Μάλλιου, Ν., Χούσος, Ν., & Καραϊσκος, Δ. (2014). Ελληνικές Επιστημονικές Δημοσιεύσεις 1998-2012: Βιβλιομετρική Ανάλυση Ελληνικών Δημοσιεύσεων σε Διεθνή Επιστημονικά Περιοδικά - Web of Science Retrieved from http://report04.metrics.ekt.gr/