ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ &
ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΩΝ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΓΙΑ ΤΗ ΣΥΝΘΕΤΙΚΗ, ΠΟΣΟΤΙΚΗ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ, ΑΠΟ ΒΙΟΛΟΓΙΚΑ ΜΕΤΑ-ΔΕΔΟΜΕΝΑ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΙΜΟΤΗΤΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΘΕΟΔΩΡΟΣ Γ. ΚΟΥΤΣΑΝΔΡΕΑΣ
ΔΙΠΛΩΜΑΤΟΥΧΟΣ ΧΗΜΙΚΟΣ ΜΗΧΑΝΙΚΟΣ ΕΜΠ

Επιβλέπων: Λ.Γ. Αλεξόπουλος
Αν. Καθηγητής ΕΜΠ

Αθήνα, Μάιος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ &
ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΩΝ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΝΟΗΜΟΣΥΝΗΣ ΓΙΑ ΤΗ ΣΥΝΘΕΤΙΚΗ, ΠΟΣΟΤΙΚΗ ΚΑΙ ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ, ΑΠΟ ΒΙΟΛΟΓΙΚΑ ΜΕΤΑ-ΔΕΔΟΜΕΝΑ ΥΨΗΛΗΣ ΔΙΑΣΤΑΣΙΜΟΤΗΤΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΘΕΟΔΩΡΟΣ Γ. ΚΟΥΤΣΑΝΔΡΕΑΣ
ΔΙΠΛΩΜΑΤΟΥΧΟΣ ΧΗΜΙΚΟΣ ΜΗΧΑΝΙΚΟΣ ΕΜΠ

ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Λ. Αλεξόπουλος, Αν. Καθ. ΕΜΠ
Α. Χατζηιωάννου, Κ. Ερευν. ΙΙΒΕΑΑ
Η. Μαγκλογιάννης, Καθ. ΠΑ.ΠΕΙ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Λ. Αλεξόπουλος, Αν. Καθ. ΕΜΠ
Α. Χατζηιωάννου, Κ. Ερευν. ΙΙΒΕΑΑ
Η. Μαγκλογιάννης, Καθ. ΠΑ.ΠΕΙ
Ν. Χονδρογιάννη, Κ.Ερευν. ΕΙΕ
Ε. Chevet, Κ. Ερευν. INSERM
Μ. Αναγνωστάκης, Αν. Καθ. ΕΜΠ
Χ. Καρανίκας, Λέκτορας, ΠΘ

Αθήνα, Μάιος 2021

Ευχαριστίες

Η παρούσα διδακτορική διατριβή εκπονήθηκε υπό την επίβλεψη του καθηγητή Λεωνίδα Αλεξόπουλου, στη Σχολή Μηχανολόγων Μηχανικών (ΕΜΠ). Θα ήθελα να τον ευχαριστήσω βαθύτατα για την ευκαιρία που μου έδωσε να πραγματοποιήσω αυτό το μακρύ ταξίδι, καθώς και για την υπομονή που έδειξε στην καθυστέρηση ολοκλήρωσής του. Ο ρόλος του ήταν καταλυτικός και χωρίς τη συνεισφορά του η διατριβή αυτή δεν θα μπορούσε να πραγματοποιηθεί. Συνεπιβλέπων υπήρξε ο ερευνητής Αριστοτέλης Χατζηιωάννου (ΙΙΒΕΑΑ). Πιστεύω πως ένα απλό ευχαριστώ δεν αρκεί για τα όσα αποκόμισα από τη συνεργασία μας. Ξεκινώντας από την περίοδο των προπτυχιακών μου σπουδών στη Σχολή Χημικών Μηχανικών (ΕΜΠ) και την εκπόνηση της διπλωματικής εργασίας υπό την επίβλεψή του, στο εργαστήριο Μεταβολικής Μηχανικής και Βιοπληροφορικής του Εθνικού Ιδρύματος Ερευνών (ΕΙΕ), η συνεργασία μας συνεχίστηκε και την περίοδο των διδακτορικών σπουδών. Όλα αυτά τα χρόνια, έχοντας το ρόλο του δασκάλου, θεωρώ πως με την καθοδήγηση και τις συμβουλές του, με βοήθησε να ωριμάσω ως ερευνητής και να αποκτήσω τη μαθηματική και κριτική σκέψη που απαιτείται για την προσέγγιση των σύγχρονων προβλημάτων στον κλάδο της βιοϊατρικής έρευνας. Οι πολύωρες συζητήσεις μας ήταν απολαυστικές και είμαι σίγουρος πως θα αποτελέσουν εφόδιο για τη μετέπειτα σταδιοδρομία μου. Θα ήθελα επίσης να τον ευχαριστήσω, γιατί πέρα απ την συνεργασία μας στο εργαστήριο, με ενέταξε στην ομάδα της εταιρείας e-NIOS Applications (e-Noesis Inspired Operational Systems) και με βοήθησε να κατανοήσω τον τρόπο με τον οποίο η σύγχρονη εφαρμοσμένη έρευνα μπορεί, και οφείλει, να οδηγήσει στο σχεδιασμό χρήσιμων καινοτόμων υπηρεσιών και προϊόντων.

Στενός συνεργάτης στο εργαστήριο Μεταβολικής Μηχανικής και Βιοπληροφορικής και στην e-NIOS υπήρξε ο Ελευθέριος Πιλάλης, ο οποίος συνεισέφερε με τις συμβουλές του στο σχεδιασμό των μεθοδολογικών προσεγγίσεων που παρουσιάζονται στη διατριβή. Θα ήθελα να τον ευχαριστήσω για την καθημερινή μας συνεργασία. Επίσης θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη του εργαστηρίου, την Όλγα Παπαδόδημα, τον Κωνσταντίνο Βουτετάκη, τον Ευθύμιο Λαδουκάκη, τον Στάθη Βλαχάβα, τη Γεωργία Κοντογιάννη και την Ιλόνα Μπινενμπάουμ, καθώς και μέλη της ομάδας της e-NIOS, τη Μαριάνθη Λογοθέτη, τον Ιωάννη Σοφιανό, τον Aitor Almanza Goikoetxea και τη Χαρά Μαστρόκαλου για τη συνεργασία και τη σχέση που χτίσαμε όλα αυτά τα χρόνια.

Στη διάρκεια της διατριβής επισκέφθηκα για έξι μήνες την ομάδα του ερευνητή Philippe Krebs στο Ινστιτούτο Παθολογίας του Πανεπιστημίου της Βέρνης (Ελβετία) και για πέντε μήνες την ομάδα του Eric Chevet στο Κέντρο για την Καταπολέμηση του Καρκίνου (Eugène Marquis) στο Πανεπιστημιακό Νοσοκομείο της Ρεν (Γαλλία). Θα ήθελα να τους ευχαριστήσω για τη φιλοξενία, για τις εποικοδομητικές συζητήσεις μας και για τη δυνατότητα που μου έδωσαν να συμμετάσχω σε ερευνητικές τους μελέτες. Οι συνεργασίες μας με βοήθησαν να αποκτήσω μια πιο ολοκληρωμένη εικόνα για τον ρόλο του μηχανικού στη βιοϊατρική έρευνα και να μάθω να επικοινωνώ με επιστήμονες διαφορετικών

ειδικοτήτων. Θα ήθελα να ευχαριστήσω τον καθηγητή Ηλία Μαγκλογιάννη (ΠΑΠΕΙ) που δέχτηκε να συμμετέχει στην τριμελή συμβουλευτική επιτροπή, καθώς και την ερευνήτρια του ΕΙΕ Νίκη Χονδρογιάννη, τον καθηγητή της Σχολής Μηχανολόγων Μηχανικών (ΕΜΠ) Μάριο Αναγνωστάκη και τον Λέκτορα του Τμήματος Πληροφορικής με Εφαρμογές στη Βιοϊατρική (ΠΘ) Χαράλαμπο Καρανίκα που δέχτηκαν να συμμετέχουν στην επταμελή εξεταστική επιτροπή.

Όλα αυτά τα χρόνια, υπήρξαν στιγμές αμφισβήτησης και ψυχολογικής κόπωσης. Είμαι ευγνώμων για τους ελάχιστους φίλους και την οικογένεια μου που ήταν πάντα εκεί, να με ενθαρρύνουν και να με στηρίζουν, υλικά και ψυχολογικά, για την ολοκλήρωση αυτού του διδακτορικού.

Περίληψη

Στην παρούσα διατριβή, αναπτύχθηκε μια νέα υπολογιστική προσέγγιση, βασισμένη στην ιδέα της ανάλυσης σημασιολογικών δικτύων, με σκοπό την ερμηνεία ομικών δεδομένων, χρησιμοποιώντας τις βιοϊατρικές οντολογίες. Οι βιοϊατρικές οντολογίες αποτελούν ελεγχόμενα λεξικά οργάνωσης της υπάρχουσας γνώσης σε κλάδους της Βιολογίας (π.χ. ασθένειες, μοριακές λειτουργίες, μεταβολικά μονοπάτια) και χρησιμοποιούνται για το σχολιασμό των βιομορίων (κυρίως γονιδίων και πρωτεϊνών), παρέχοντας μια πολύπλευρη σημασιολογική περιγραφή για το κάθε ένα. Για την υλοποίησή της συγκεκριμένης προσέγγισης, κατασκευάστηκε μια νέα βιβλιοθήκη λογισμικού. Η βιβλιοθήκη περιέχει συναρτήσεις, κλάσεις και ροές εργασιών για την ενσωμάτωση, επεξεργασία και χρήση των οντολογιών, καθώς και την τοπολογική ανάλυση στους γράφους αυτών. Για την ερμηνεία των ομικών δεδομένων κατασκευάστηκε μια μη εποπτευόμενη, αυτοματοποιημένη, αναλυτική ροή εργασιών, η οποία συνδυάζει τις μεθόδους της ανάλυσης μονοπατιών και της ιεράρχησης γονιδίων, αξιοποιώντας το σχολιασμό και τη δομή των βιοϊατρικών οντολογιών. Σκοπός της είναι η μετατροπή της κατανομής των σημάτων ενός πειράματος υψηλής απόδοσης σε δίκτυο σημασιολογικών όρων και ο εντοπισμός των κεντρικών ρυθμιστικών βιομορίων πάνω σε αυτό. Αυτή η ροή εργασιών ονομάζεται BioInfoMiner και ενσωματώνει αλγοριθμικές ενότητες που έχουν αναπτυχθεί ως ξεχωριστά εργαλεία στο παρελθόν. Τα εργαλεία αυτά επανασχεδιάστηκαν με σκοπό να ενσωματωθούν σε ένα ενοποιημένο υπολογιστικό εργαλείο, ικανό να λειτουργήσει σε συστήματα τεχνολογίας νέφους. Επιπλέον, δημιουργήθηκε μια ροή εργασιών διαδραστικής απεικόνισης των αποτελεσμάτων σε περιβάλλον διεπαφής. Μια έκδοση του BioInfoMiner ενσωματώθηκε σε διαδικτυακή εφαρμογή, βασισμένη στο υπολογιστικό σύστημα Galaxy, για την ανάλυση μεταγενωμικών δεδομένων. Για τη συγκριτική ανάλυση σημασιολογικών δικτύων αναπτύχθηκαν κατάλληλες παραμετροποιήσιμες συναρτήσεις, ενώ πραγματοποιήθηκε ποιοτική σύγκριση των υπαρχόντων σημασιολογικών μέτρων. Στα πλαίσια της παρούσας διατριβής, ο BioInfoMiner και η εν λόγω βιβλιοθήκη χρησιμοποιήθηκαν σε τρεις μελέτες. Η πρώτη αφορούσε στην ερμηνεία μεταγραφωμικών δεδομένων, σχετικά με τη μελέτη του ρόλου της πρωτεΐνης TRAIL στη λειτουργία των NK κυττάρων σε συνθήκες ιογενούς λοίμωξης. Στη δεύτερη μελέτη, κατασκευάστηκε το σημασιολογικό προφίλ του μηχανισμού της πρωτεόστασης για εκατοντάδες οργανισμούς, με στόχο την αξιολόγηση της εξελικτικής του αποτύπωσης. Τέλος μελετήθηκε το πρωτεϊνικό δίκτυο που αλληλεπιδρά με τις πρωτεΐνες του ιού SARS-CoV-2, ώστε να εντοπιστούν τα βασικά μοτίβα της παθογόνου δράσης του και η συσχέτιση του με αντίστοιχα δίκτυα άλλων παθογόνων ιών. Οι παραπάνω μελέτες οδήγησαν σε νέα ευρήματα αξίας, υποστηρίζοντας την αποτελεσματικότητα του BioInfoMiner, σχετικά με τη μείωση της διαστασιμότητας και πολυπλοκότητας των συνόλων ομικών δεδομένων. Επίσης, επιβεβαίωσαν ότι η εν λόγω συστημική σημασιολογική προσέγγιση μπορεί να αναδείξει τα κομβικά κυτταρικά γεγονότα και ρυθμιστικά βιομόρια σε μια υπό μελέτη φαινοτυπική συνθήκη και να εξαγάγει κρίσιμη, νέα βιολογική πληροφορία.

Εκτεταμένη Περίληψη

Εισαγωγή

Ο όρος «ομικά δεδομένα» χρησιμοποιείται για να προσδιορίσει τα ψηφιοποιημένα δεδομένα που προκύπτουν από βιολογικά πειράματα υψηλής απόδοσης και περιγράφουν τη συνολική κατάσταση ενός συγκεκριμένου τύπου κυτταρικών βιομορίων και των τροποποιήσεών τους (π.χ. γενωμικά, επιγενωμικά, πρωτεωμικά, κ.α. δεδομένα), υπό συγκεκριμένες συνθήκες. Οι τεχνολογίες υψηλής απόδοσης στη Μοριακή Βιολογία άρχισαν να αναπτύσσονται και να εφαρμόζονται μαζικά, από τις αρχές του 21ου αιώνα, κυρίως μετά την αποκρυπτογράφηση του ανθρώπινου γονιδιώματος, οπότε και έγινε αντιληπτή η εξαιρετική πολυπλοκότητα της γονιδιακής ρύθμισης και κυτταρικής λειτουργίας. Πλέον, χρησιμοποιούνται για τη συστημική μελέτη πολύπλοκων ασθενειών και σύνθετων βιολογικών ερωτημάτων, ποσοτικοποιώντας τα επίπεδα έκφρασης, ρύθμισης και λειτουργίας των βιομορίων μέσα στο κύτταρο. Η παράλληλη ανάπτυξη των υπολογιστικών συστημάτων και των αναλυτικών μεθόδων για την αυτοματοποιημένη ανάλυση των ομικών δεδομένων, οδήγησε στη εξαγωγή κρίσιμης γνώσης για χιλιάδες βιομόρια, σε διαφορετικές φαινοτυπικές συνθήκες, βιολογικά συστήματα και οργανισμούς. Για την αποτελεσματική οργάνωση και επαναχρησιμοποίηση αυτής της γνώσης, κατασκευάστηκαν κατάλληλες βάσεις δεδομένων. Μια κατηγορία αυτών είναι οι βιοϊατρικές οντολογίες, οι οποίες αποτελούν ελεγχόμενα λεξικά με ιεραρχική δομή, όπου περιγράφεται η υπάρχουσα γνώση σε έναν συγκεκριμένο κλάδο της Βιολογίας σε μορφή αναγνώσιμη από τους υπολογιστές.

Στην επιστήμη της Πληροφορικής και της Τεχνητής Νοημοσύνης, ο όρος οντολογία αναφέρεται στον τυπικό και ρητό ορισμό μιας εννοιολογικής αναπαράστασης, για την ολιστική περιγραφή ενός συγκεκριμένου τομέα. Η αναπαράσταση αυτή αποτελείται κυρίως από στιγμιότυπα συγκεκριμένων κλάσεων (δηλ. σημασιολογικούς όρους), και τις μεταξύ τους σχέσεις, σχηματίζοντας συνήθως ένα ελεγχόμενο λεξικό με δομή άκυκλου κατευθυνόμενου γράφου. Ξεκινώντας από την κορυφή του γράφου, ο γενικός όρος διακλαδίζεται σε ειδικότερους, καταλήγοντας σταδιακά στα φύλλα, δηλαδή τους πιο εξειδικευμένους εννοιολογικά όρους. Στον κλάδο της Βιοϊατρικής έχουν κατασκευαστεί αρκετές οντολογίες που καλύπτουν τομείς όπως η κυτταρική λειτουργία, οι ασθένειες του ανθρώπου, η ανατομία και οι αναπτυξιακές διαδικασίες οργανισμών-μοντέλων κ.α. Η ευρεία χρήση τους οφείλεται αφενός στο οτι χρησιμεύουν ως μέσο οργάνωσης της υπάρχουσας γνώσης και αφετέρου στο ότι αποτελούν λεξικά για το σχολιασμό των βιομορίων (κυρίως γονιδίων και πρωτεϊνών), σχηματίζοντας μια πολύπλευρη σημασιολογική περιγραφή για το κάθε ένα. Οι σχολιασμοί αυτοί επιτρέπουν τη λειτουργική ερμηνεία των ομικών δεδομένων, καθώς και την ενοποιημένη ανάλυση ετερογενών δεδομένων. Για το λόγο αυτό, παράλληλα με την κατασκευή των οντολογιών, αναπτύχθηκαν διάφορες αναλυτικές μεθοδολογίες για τη μετα-ανάλυση των ομικών δεδομένων, αξιοποιώντας τον γενωμικό και πρωτεωμικό σχολιασμό των οντολογιών και τη δομή των σημασιολογικών τους γράφων.

Γενικά, η επεξεργασία των ομικών δεδομένων καταλήγει σε μια λίστα διαφοροποιημένων γονιδίων, RNAs ή πρωτεϊνών ανάμεσα στις κατηγορίες των εξεταζόμενων δειγμάτων. Συνήθως οι λίστες αυτές αναφέρονται ως "λίστες γονιδίων", καθώς η ροή της γενετικής πληροφορίας για την παραγωγή όλων των βιομορίων ξεκινά από τα γονίδια, δηλαδή όλα τα βιομόρια προέρχονται από ένα ή περισσότερα γονίδια. Η πιο ευρέως χρησιμοποιούμενη μεθοδολογία για τη μετα-ανάλυση αυτών των λιστών, είναι η ανάλυση μονοπατιών, η οποία αξιοποιεί τον οντολογικό σχολιασμό των βιομορίων, ώστε να εξετάσει τις συνεργιστικές τους σχέσεις και να αναδείξει το διαφοροποιημένο δίκτυο μηχανισμών στο οποίο εμπλέκονται. Η ανάλυση πραγματοποιείται με διάφορα στατιστικά τεστ, τα οποία εντοπίζουν τους υπεραντιπροσωπευμένους (ή εμπλουτισμένους) σημασιολογικούς όρους για μια λίστα βιομορίων. Μια άλλη μεθοδολογία που χρησιμοποιείται για τον εντοπισμό των πιο σημαντικών βιομορίων, για μια δεδομένη συνθήκη (π.χ. ασθένεια, μοριακό μηχανισμό), είναι η ιεράρχηση γονιδίων. Η αξιολόγηση των υποψήφιων γονιδίων βασίζεται στην συσχέτισή τους είτε με γονίδια αναφοράς είτε με λέξεις κλειδιά και σημασιολογικούς όρους, που θεωρούνται αντιπροσωπευτικά για την εξεταζόμενη συνθήκη. Για την ιεράρχηση τους σε σχέση με τα κριτήρια αναφοράς, χρησιμοποιείται ο οντολογικός σχολιασμός τους και λοιπά δεδομένα από πληθώρα βάσεων.

Συνολικά, ο οντολογικός σχολιασμός περιγράφει τα λειτουργικά χαρακτηριστικά και το βιολογικό ρόλο ενός βιομορίου. Επομένως περιέχει την απαιτούμενη πληροφορία για να συγκριθούν δύο ή περισσότερα βιομόρια, με βάση τις λειτουργίες τους, τους μηχανισμούς που εμπλέκονται ή τις ασθένειες με τις οποίες συσχετίζονται. Η σύγκριση αυτή ονομάζεται σημασιολογική, καθώς εξετάζει την ομοιότητα ομάδων σημασιολογικών όρων, αξιοποιώντας την τοπολογία τους πάνω στο γράφο της εκάστοτε οντολογίας. Η σημασιολογική ανάλυση προέρχεται από την επιστήμη της Γλωσσολογίας, όπου και χρησιμοποιείται για τον υπολογισμό της ομοιότητας λέξεων και φράσεων. Στην Βιοϊατρική, έχουν προταθεί διάφορες τεχνικές και μέτρα σημασιολογικής ανάλυσης, με στόχο την ακριβή εκτίμηση της ομοιότητας των όρων και κατ΄επέκταση των βιομορίων.

Τεχνικό Μέρος

Στη συγκεκριμένη διδακτορική διατριβή, σχεδιάστηκε μια νέα υπολογιστική προσέγγιση, βασισμένη στην ιδέα της ανάλυσης σημασιολογικών δικτύων, με σκοπό τη χρήση των βιοϊατρικών οντολογιών για την πολύπλευρη ερμηνεία ομικών δεδομένων. Για την υλοποίηση κατασκευάστηκε μια νέα βιβλιοθήκη λογισμικού σε γλώσσα προγραμματισμού Python 2.7, η οποία περιέχει συναρτήσεις, κλάσεις και ροές εργασιών για την επεξεργασία και χρήση των οντολογιών και των γενωμικών σχολιασμών τους, την αποθήκευσή τους σε κατάλληλες συλλογές της MongoDB, την τοπολογική ανάλυση στους γράφους τους και τη σημασιολογική σύγκριση βιομορίων. Η εν λόγω υπολογιστική προσέγγιση υλοποιήθηκε ως μια μη εποπτευόμενη, αυτοματοποιημένη, αναλυτική ροή εργασιών, που ονομάζεται BioInfoMiner, και συνδυάζει τις μεθόδους της ανάλυσης μονοπατιών και της ιεράρχησης γονιδίων, με σκοπό να εντοπίσει το κρίσιμο δίκτυο σημασιολογικων όρων που συσχετίζεται με το αποτέλεσμα της

ομικής ανάλυσης και τα βιομόρια εκείνα, που έχουν κεντρικό ρυθμιστικό ρόλο πάνω στο δίκτυο αυτό. Για την υλοποίηση επανασχεδιάστηκαν δύο αλγόριθμοι (StRAnGER, GOrevenge) που είχαν αναπτυχθεί στο παρελθόν στο εργαστήριο Μεταβολικής Μηχανικής και Βιοπληροφορικής του Εθνικού Ιδρύματος Ερευνών (ΕΙΕ, Αθήνα, Ελλάδα). Η νέα υλοποίηση επιτρέπει την εύκολη ενσωμάτωση νέων οντολογιών, για οποιονδήποτε οργανισμό, σε αντίθεση με τις προηγούμενες εκδόσεις που υποστηρίζουν μόνο την ανάλυση με οντολογικούς σχολιασμούς του ανθρώπου και του ποντικού. Επιπλέον, ορισμένα βήματα της ανάλυσης του StRAnGER παραλληλοποιήθηκαν για την ταχύτερη εκτέλεση τους. Όσον αφορά τον GOrevenge, τροποποιήθηκε ο αλγόριθμος και τα κριτήρια ιεράρχησης των γονιδίων πάνω στο σημασιολογικό δίκτυο, με σκοπό την εξαγωγή μιας βέλτιστης, μοναδικά ορισμένης υπογραφής. Η συνολική αρχιτεκτονική της ροής εργασιών προβλέπει την εύκολη ενσωμάτωσή της σε εικονικές υπολογιστικές μηχανές, στο πλαίσιο των τεχνολογιών νέφους. Επιπλέον δημιουργήθηκε μια ροή εργασιών απεικόνισης των αποτελεσμάτων σε μορφή διμερούς γράφου, όπου στην μια πλευρά του αποτυπώνεται το κρίσιμο σημασιολογικό δίκτυο και στην άλλη τα ιεραρχημένα γονίδια. Το τελικό διάγραμμα προβάλλεται σε περιβάλλον διεπαφής, με χρήση της γλώσσας προγραμματισμού JavaScript και της βιβλιοθήκης D3, η οποία επιτρέπει τη διαδραστική απεικόνιση.

Η νέα βιβλιοθήκη λογισμικού επιτρέπει την αναδιαμόρφωση της ροής εργασιών του BioInfoMiner, ώστε να εφαρμοστεί σε ένα ευρύ φάσμα εργασιών βιοπληροφορικής ανάλυσης, που εκτελούνται σε διαφορετικά λειτουργικά συστήματα και περιβάλλοντα. Για παράδειγμα, στα πλαίσια της διατριβής, μια έκδοση του BioInfoMiner ενσωματώθηκε σε διαδικτυακή εφαρμογή, βασισμένη στο υπολογιστικό σύστημα Galaxy, την πλατφόρμα ANASTASIA, για τη λειτουργική ανάλυση μεταγενωμικών δειγμάτων. Ο επικυρωμένος σχολιασμός της Gene Ontology, σύμφωνα με τη βάση δεδομένων UniProt/SwissProt, για όλους τους προκαρυωτικούς οργανισμούς, ενσωματώθηκε σε ένα ενιαίο σχήμα, όπου κάθε γονίδιο χαρακτηρίζεται από ένα μοναδικό σετ σημασιολογικών όρων, ανεξάρτητα από τους οργανισμούς στους οποίους συναντάται. Επιπλέον η ροή εργασιών του BioInfoMiner τροποποιήθηκε ώστε να ενσωματωθεί ως ξεχωριστό αναλυτικό βήμα στις σειρές εργασιών της πλατφόρμας ANASTASIA, δεχόμενο ως είσοδο το αποτέλεσμα της ανάλυσης του εργαλείου BLASTp, είτε απευθείας είτε μετά από μια διαδικασία φιλτραρίσματος των αποτελεσμάτων με κριτήρια ορισμένα από τον χρήστη. Η συγκεκριμένη υλοποίηση καταδεικνύει την ευελιξία και ευχρηστότητα του BioInfoMiner, που μπορεί να ενσωματωθεί εύκολα, ως ενοποιημένο και αυτοματοποιημένο εργαλείο σε διάφορα υπολογιστικά περιβάλλοντα.

Εφαρμογές

Η ιδέα της σημασιολογικής σύγκρισης των βιομορίων με βάση τον οντολογικό σχολιασμό τους, έχει οδηγήσει στη μελέτη της δομής των αντίστοιχων γράφων, με αποτέλεσμα τον ορισμό τοπολογικών ιδιοτήτων και την ανάπτυξη σημασιολογικών μέτρων, για την εκτίμηση της ομοιότητας δύο όρων. Στην παρούσα διατριβή αξιολογήθηκαν οι υπάρχουσες τεχνικές υπολογισμού της σημασιολο-

γικής ομοιότητας, με βάση ένα προτεινόμενο πλαίσιο κανόνων σύμφωνο με την ανθρώπινη λογική. Για τον σκοπό αυτό κατασκευάστηκε ένας τεχνητός οντολογικός γράφος και χρησιμοποιώντας το εν λόγω πλαίσιο, ορίστηκαν ζεύγη όρων με σημαντικές σημασιολογικές διαφορές. Στη συνέχεια αξιολογήθηκαν διάφορα μέτρα για την ικανότητά τους να αναπαράγουν την προτεινόμενη ιεράρχηση των ζευγών, βάσει των σημασιολογικών ομοιοτήτων. Η ανάλυση κατέδειξε πως τα αποτελεσματικότερα μέτρα, χωρίζονται σε δύο κατηγορίες λόγω των ελλείψεών τους: στα αυστηρά μέτρα που υπολογίζουν ίδιες τιμές ομοιότητας για ανόμοια ζεύγη και στα μέτρα που παράγουν τιμές μεγαλύτερης διασποράς και είναι επιρρεπή σε λανθασμένες εκτιμήσεις στην ιεράρχηση των ζευγών. Καθώς η επιλογή του κατάλληλου μέτρου αποτελεί ένα πρόβλημα αντιστάθμισης της μεροληψίας πάνω στον οντολογικό γράφο, είναι αναγκαία η ανάπτυξη νέων, βελτιωμένων μέτρων, που θα υιοθετούν το προτεινόμενο πλαίσιο κανόνων στο σύνολό του.

Στα πλαίσια της παρούσας διατριβής, ο BioInfoMiner και άλλες ροές εργασιών και ρουτίνες της βιβλιοθήκης χρησιμοποιήθηκαν σε τρεις μελέτες. Η πρώτη αφορούσε τη σημασιολογική ερμηνεία μεταγραφωμικών δεδομένων, που παρήχθησαν από πειράματα υψηλής απόδοσης για τη μελέτη του μη κανονικού ρόλου της πρωτεΐνης TRAIL στην ενεργοποίηση των NK λεμφοκυττάρων σε συνθήκες ιογενούς λοίμωξης. Τα NK κύτταρα έχουν σημαντικό ρόλο στην εγγενή ανοσία του ανθρώπινου οργανισμού κατά την ιογενή λοίμωξη και το σχηματισμό όγκων. Η δράση τους περιλαμβάνει την έκκριση είτε κυτταροτοξικών κοκκιοκυττάρων είτε κυτταροκινών με στόχο τη νέκρωση των κυττάρων-στόχων. Μια απο τις παραγόμενες κυτταροκίνες είναι και η πρωτεΐνη TRAIL, που προσδένεται σε συγκεκριμένους υποδοχείς των κυττάρων-στόχων και ενεργοποιεί τον μηχανισμό της απόπτωσης. Πολλές μελέτες έχουν καταγράψει μη κανονικούς ρόλους της TRAIL, κατά τους οποίους όχι μόνο δεν ενεργοποιείται η απόπτωση, αλλά παρεμποδίζεται ο κυτταρικός θάνατος και η γενικότερη αντιμετώπιση της λοίμωξης, ενώ σε περιπτώσεις όγκων ενισχύονται οι μηχανισμοί της αναπαραγωγής, εισβολής και μετάστασης. Στην έρευνα που πραγματοποιήθηκε στο Ινστιτούτο Παθολογίας του Πανεπιστημίου της Βέρνης (Ελβετία), πραγματοποιήθηκε σειρά πειραμάτων για τη μελέτη του μη κανονικού ρόλου της TRAIL στην λειτουργία των NK κυττάρων ποντικών, κατά τη λοίμωξη από το λεμφοκυτταρικό ιό της χοριομηνιγγίτιδας (LCMV). Τα αποτελέσματα έδειξαν πως η παραγωγή της TRAIL από τα NK κύτταρα σχετίζεται με τη ρύθμιση της δράσης των CD8 T κυττάρων, εμποδίζοντας την ενεργοποίησή τους και καθυστερώντας την αντιμετώπιση της λοίμωξης. Επιπλέον, η TRAIL εμποδίζει την παραγωγή άλλων κυτταροκινών, όπως η ιντερφερόνη γάμμα (IFN-γ), και επάγει την παραγωγή του γρανενζύμου Β μέσω του μονοπατιού της ιντερλευκίνης 15 (IL-15), πιθανότατα συμμετέχοντας στην ενεργοποίηση του μονοπατιού PI3K-AKT-mTOR. Με σκοπό την περαιτέρω μελέτη του ρόλου της TRAIL, πραγματοποιήθηκε μεταγραφωμική ανάλυση σε δεκατέσσερα δείγματα NK κυττάρων, τα οποία απομονώθηκαν από τη σπλήνα των ποντικών και διαχωρίστηκαν σε τέσσερις κατηγορίες με βάση τη μοριακή και παθολογική τους κατάσταση: στελέχη άγριου-τύπου (Wild Type - WT) και στελέχη με έλλειψη

της TRAIL (TRAIL Knock Out - KO), υγιή και μολυσμένα με τον ιό LCMV. Η επεξεργασία των δεδομένων περιελάμβανε τη συναρμολόγηση του γονιδιώματος, την ποσοτικοποίηση της έκφρασης των γονιδίων σε κάθε δείγμα και τον εντοπισμό των διαφορικά εκφρασμένων γονιδίων μεταξύ των μολυσμένων και υγιών δειγμάτων της ίδια μοριακής κατάστασης. Επομένως, οι δύο λίστες γονιδίων που προέκυψαν αποτύπωσαν τη διαφοροποίηση της μεταγραφικής διαδικασίας στα NK κύτταρα κατά τη λοίμωξη στην περίπτωση των στελεχών άγριου-τύπου, και σε αυτή των στελεχών με έλλειψη της TRAIL. Στη συνέχεια χρησιμοποιήθηκε ο BioInfoMiner και οι οντολογίες Gene Ontology και Reactome για τη λειτουργική ερμηνεία αυτών των λιστών. Επιπλέον με χρήση ορισμένων συναρτήσεων της βιβλιοθήκης λογισμικού, συγκρίθηκαν οι τοπολογίες των δύο παραγόμενων γράφων εμπλουτισμένων όρων και προσδιορίστηκαν οι μοναδικά εμπλουτισμένοι σε μια από τις δύο μοριακές καταστάσεις. Με τον τρόπο αυτό εντοπίστηκαν βιολογικές διαδικασίες και μοριακά μονοπάτια που διαφοροποιήθηκαν μόνο σε μια από τις δύο καταστάσεις λοίμωξης, υποδεικνύοντας την τροποποίηση της λειτουργίας των NK κυττάρων κατά την ενεργοποίησή τους, λόγω έλλειψης της TRAIL. Τα αποτελέσματα της ανάλυσης επαλήθευσαν ορισμένα από τα προηγούμενα πειραματικά ευρήματα για τον μη κανονικό ρόλο της TRAIL. Συγκεκριμένα, η ανοσολογική απόκριση άγριου-τύπου συσχετίστηκε με όρους που αφορούν την αρνητική ρύθμιση της παραγωγής κυτταροκινών και τα σηματοδοτικά μονοπάτια ιντερλευκινών και PI3K-AKT-mTOR. Αντίθετα, σε συνθήκες έλλειψης της TRAIL εμπλουτίστηκαν όροι σχετικοί με τη θετική ρύθμιση της παραγωγής κυτταροκινών και την αρνητική ρύθμιση του δικτύου PI3K/AKT. Οι λίστες των ιεραρχημένων γονιδίων παρουσίασαν μεγάλη επικάλυψη, με πολλά κοινά γονίδια που παράγουν σημαντικές πρωτεΐνες για την ανοσολογική απόκριση (ιντερλευκίνες, ιντερφερόνες, χημειοκίνες και παράγοντες νέκρωσης όγκων) και στις δύο περιπτώσεις. Συμπερασματικά, η ερμηνεία των μεταγραφωμικών δεδομένων κατάφερε να αναδείξει τις μικροδιαφορές στην ανοσολογική απόκριση των δύο μοριακών συνθηκών, επαληθεύοντας ευρήματα των προηγούμενων πειραμάτων.

Στη δεύτερη μελέτη, κατασκευάστηκε η σημασιολογική περιγραφή του μηχανισμού της πρωτεϊνικής ομοιόστασης (πρωτεόσταση) σε εκατοντάδες οργανισμούς, με στόχο την αξιολόγηση της εξελικτικής της αποτύπωσης, κυρίως στις βασικές ταξινομικές βαθμίδες (Ευκαρυώτες, Βακτήρια και Αρχαία). Η πρωτεόσταση συναντάται σε όλα τα είδη κυττάρων και αποτελείται από ένα πολύπλοκο δίκτυο διαδικασιών που επηρεάζει τα επίπεδα έκφρασης, τη διαμόρφωση και τη συσσώρευση των πρωτεϊνών, ελέγχοντας τη σύνθεση, την αναδίπλωση, τη μεταφορά και την αποδόμηση τους. Εξελικτικά, ο μηχανισμός της πρωτεόστασης έχει διαμορφωθεί για κάθε οργανισμό, ώστε να προσαρμοστεί η κυτταρική λειτουργία στις εγγενείς και εξωγενείς απαιτήσεις, ενώ η μη φυσιολογική λειτουργία της συσχετίζεται με την κυτταρική γήρανση και την εμφάνιση ασθενειών. Οι παραδοσιακές τεχνικές φυλογενετικής ανάλυσης βασίζονται κυρίως στη σύγκριση συγκεκριμένων γονιδιακών και πρωτεϊνικών αλληλουχιών, που υπάρχουν σε όλα τα εξεταζόμενα είδη, και καταλήγουν στην κατασκευή ενός φυλογενετικού δέντρου, όπου είδη με όμοιες αλληλουχίες

ταξινομούνται στο ίδιο κλαδί. Στη συγκεκριμένη μελέτη, παρουσιάστηκε μια νέα μεθοδολογία φυλογενετικής ανάλυσης, η οποία βασίζεται στο λειτουργικό προφίλ των οργανισμών και όχι σε μεμονωμένες μοριακές αλληλουχίες. Συγκεκριμένα, σχηματίστηκε το σημασιολογικό δίκτυο της πρωτεόστασης, με όρους της Gene Ontology (Biological Process), για εκατοντάδες οργανισμούς (93 Αρχαία, 250 Βακτήρια και 94 Ευκαρυώτες), αναλύοντας λίστες γονιδίων που εμπλέκονται σε κεντρικούς μηχανισμούς της, σύμφωνα με τη βιβλιογραφία. Στη συνέχεια πραγματοποιήθηκε σημασιολογική σύγκριση αυτών των δικτύων, ώστε να κατασκευαστεί το αντίστοιχο φυλογενετικό δέντρο και να ελεγχθεί η ικανότητα της πρωτεόστασης να λειτουργήσει ως δείκτης της εξέλιξης των ειδών. Επιπλέον εντοπίστηκαν τα τμήματα (μηχανισμοί) των δικτύων που εμφανίζονται σε όλα τα εξεταζόμενα είδη, αλλά και αυτά που αποτελούν κριτήριο διαχωρισμού των βασικών ταξινομικών βαθμίδων. Τέλος, μελετήθηκε η επίδραση της πρωτεόστασης στην ικανότητα άλλων κυτταρικών μηχανισμών να χρησιμοποιούνται ως εξελικτικοί δείκτες. Τα αποτελέσματα της φυλογενετικής ανάλυσης έδειξαν πως η πρωτεόσταση μπορεί να διαχωρίσει τα τρια ταξινομικά βασίλεια με ακρίβεια παρόμοια με αυτή του ριβοσωμικού RNA (rRNA), το οποίο αποτελεί το πιο ευρέως χρησιμοποιούμενο κριτήριο φυλογενετικής σύγκρισης. Επιπλέον αποδείχτηκε πιο αποτελεσματική από τις αλληλουχίες των πρωτεϊνών θερμικού σοκ (HSP40 και HSP70), οι οποίες είναι κεντρικά μόρια για το συνολικό μηχανισμό της πρωτεόστασης. Τα κοινά τμήματα της πρωτεόστασης που συναντώνται σε όλα τα είδη αφορούν στην αναδίπλωση, μεταφορά και μεταβολισμό των πρωτεϊνών στο κύτταρο. Αντιθέτως, όσο ανώτερος εξελικτικά είναι ένας οργανισμός, τόσο εμφανίζονται περισσότερες ρυθμιστικές διαδικασίες που εμπλέκονται με την οργάνωση, τη σηματοδότηση, την απόκριση και το θάνατο των κυττάρων. Στο τελευταίο μέρος της ανάλυσης εκτιμήθηκε η εξελικτική αποτύπωση άλλων συντηρημένων βιολογικών μηχανισμών, των οποίων η ακρίβεια στον διαχωρισμό των ταξινομικών βασιλείων μειώνεται, όταν αφαιρούνται τα τμήματα της πρωτεόστασης απ το σημασιολογικό τους προφίλ. Η συγκεκριμένη μελέτη κατέδειξε την πολυλειτουργικότητα της νέας βιβλιοθήκης λογισμικού, καθώς και την ευελιξία του BioInfoMiner, στην ενσωμάτωση και σημασιολογική ερμηνεία δεδομένων για εκατοντάδες οργανισμούς, με διαφορετικό οντολογικό σχολιασμό. Η προτεινόμενη μεθοδολογία φυλογενετικής ανάλυσης διαφέρει από τις υπάρχουσες καθώς αξιοποιεί τα λειτουργικά χαρακτηριστικά των βιομορίων και όχι τις μεμονωμένες αλληλουχίες τους, δίνοντας σημαντικές πληροφορίες για το πως η εξελικτική πίεση οδήγησε στη δημιουργία νέων μηχανισμών και πιο πολύπλοκων συστημάτων. Η εφαρμογή της έδειξε πως ο μηχανισμός της πρωτεόστασης συνδέεται με άλλους κεντρικούς μηχανισμούς της κυτταρικής λειτουργίας και μπορεί να θεωρηθεί ως δείκτης της πολυπλοκότητας της.

Στην τελευταία εργασία, με αφορμή την παγκόσμια πανδημία του SARS-CoV-2, μελετήθηκε το πρωτεϊνικό δίκτυο των ανθρώπινων κυττάρων που αλληλεπιδρά με τις πρωτεΐνες του SARS-CoV-2 στα διάφορα στάδια του ικού κύκλου ζωής, προκειμένου να εντοπιστούν τα βασικά μοτίβα της παθογόνου δράσης του και η επίδραση τους στη λειτουργία των ανθρώπινων κυττάρων. Επιπλέον πραγματοποιήθηκε συγκριτική λειτουργική ανάλυση του εν λόγω

πρωτεϊνικού δικτύου με αντίστοιχα δίκτυα άλλων παθογόνων ιών, χρησιμοποιώντας γνωστές βιοϊατρικές οντολογίες, με σκοπό να εντοπιστεί πιθανή συσχέτιση αυτών των ιών με τον SARS-CoV-2. Τα πρωτεϊνικά δίκτυα προσδιορίστηκαν από παρόμοια πειράματα με τη μέθοδο σύζευξης χρωματογραφίας συγγένειας με φασματομετρίας μάζας (AP-MS), στα οποία ανιχνεύτηκαν οι πρωτεΐνες ανθρώπινων κυττάρων που αντιδρούν άμεσα με πρωτεΐνες του εκάστοτε ιού. Η σημασιολογική ερμηνεία του δικτύου του SARS-CoV-2 με τις οντολογίες Gene Ontology και Reactome ανέδειξε πως ο ιός αντιδρά με πληθώρα πρωτεϊνών που συμμετέχουν στην οργάνωση των κυτταρικών τμημάτων και της μεμβράνης, στη ρύθμιση του καταβολισμού και της σταθερότητας των RNA μορίων και σε διαδικασίες της πρωτεόστασης, όπως η αναδίπλωση και η μεταφορά των πρωτεϊνών και η απόκριση στο στρες του ενδοπλασματικού δικτύου. Επιπρόσθετα, η ερμηνεία με τη χρήση της MGIMP ανέδειξε όρους σχετικούς με διάφορες ασθένειες και προβλήματα του κυκλοφορικού συστήματος και νευροεκφυλιστικές παθήσεις. Η ιεράρχηση των πρωτεϊνών κατέληξε σε μια διευρυμένη λίστα ογδόντα πρωτεϊνών, μειώνοντας την αρχική λίστα περίπου κατά τέσσερις φορές, συμπεριλαμβάνοντας πρωτεΐνες με κομβικό ρόλο στα σημασιολογικά δίκτυα που κατασκευάστηκαν, ενώ πολλές εξ αυτών μελετώνται ως στόχοι για τη θεραπεία της COVID-19. Η συγκριτική ανάλυση με τα πρωτεϊνικά δίκτυα των άλλων ιών φανέρωσε υψηλή συσχέτιση του SARS-CoV-2 με τους εντεροϊούς Coxsackievirus A10 και Rhinovirus C15. Συνολικά, η μελέτη επαλήθευσε την υπάρχουσα γνώση για τη δράση και τις επιπτώσεις του SARS-CoV-2 στη λειτουργία των κυττάρων, ενώ τόσο η ιεραρχημένη λίστα πρωτεϊνών, όσο και ο λειτουργικός και φαινοτυπικός συσχετισμός του με τους εντεροϊούς μπορούν να οδηγήσουν σε περαιτέρω πειράματα και εξαγωγή νέας γνώσης.

Επίλογος

Οι μεθοδολογίες που αναπτύχθηκαν στην παρούσα διατριβή, καθώς και οι εφαρμογές τους σε διαφορετικές μελέτες, απέδειξαν την πολυλειτουργικότητα της νέας βιβλιοθήκης λογισμικού για την ανάλυση των σημασιολογικών δικτύων. Γενικά, η εξαγωγή πληροφορίας/γνώσης με τη χρήση σημασιολογικών δικτύων βασίζεται κυρίως στην αξιοποίηση της υπάρχουσας γνώσης και δεν επηρεάζεται από τις τιμές των πειραματικών δεδομένων. Επομένως η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί για οποιοδήποτε τύπο βιομορίων υπάρχει οντολογικός σχολιασμός. Αντίστοιχα, η ροή εργασιών του BioInfoMiner, που αποτελεί ένα ενοποιημένο εργαλείο, ικανό να εκτελεστεί σε διαφορετικά υπολογιστικά συστήματα, μπορεί να χρησιμοποιηθεί για την βιολογική ερμηνεία διαφορετικών ομικών δεδομένων με μια πληθώρα οντολογιών. Τόσο οι τεχνολογίες υψηλής απόδοσης, όσο και ο τρόπος οργάνωσης της γνώσης στα οντολογικά σχήματα, έχουν φτάσει σε ένα επίπεδο ωριμότητας και έχουν καταστεί αναπόσπαστα εργαλεία για την επιστήμη της Βιολογίας Συστημάτων. Για τον λόγο αυτό, οι προτεινόμενες μεθοδολογίες για τη συστημική σημασιολογική ερμηνεία και σύγκριση των ομικών δεδομένων είναι χρήσιμες όχι μόνο για το σήμερα της βιοϊατρικής έρευνας αλλά και για το μέλλον.

Abstract

In this doctoral study, a novel computational approach was devised for the interpretation of various omics data, based on the semantic network analysis, exploiting the biomedical ontologies. Biomedical ontologies constitute controlled vocabularies, which are used to organise the existing knowledge in a specific biological domain and annotate the biomolecules (mainly genes and proteins), providing a multi-faceted semantic description for each one. A novel software library was constructed in order to implement the above computational approach. The library contains versatile classes and configured workflows for the integration, processing and manipulation of biomedical ontologies, as well as the topological analysis on their graphs. Within the frames of that library, an unsupervised, automated, analytical workflow was developed for the semantic interpretation of omics data, which combines the execution of pathway analysis and gene prioritization, using the annotation and structure of biomedical ontologies. Its final goal is to transform the distribution of signals of a high-throughput experiment into a semantic network and detect the pivotal regulatory biomolecules into it. The workflow, named BioInfoMiner, integrates algorithmic modules that have been developed as separate tools in the past. These tools were redesigned so as to be optimally integrated in a unified analytical pipeline, operating in virtual machines of cloud technologies. Besides, a novel workflow was constructed for the interactive visualization of the results in front-end interfaces. An instantiation of BioInfoMiner was integrated in a Galaxy-based platform for the analysis of metagenomic data. Additional modules were developed for the comparative analysis of semantic networks, while a qualitative evaluation of semantic similarity measures was performed. Within the frames of this thesis, BioInfoMiner and modules of the software library were used in three different projects. The first project was related to the semantic interpretation of transcriptomic data, derived from the study of the role of TRAIL in the activation of NK cells, during viral infection. In the second, the semantic profile of proteostasis machinery for thousands of species was calculated, targeting to evaluate its evolutionary imprinting. Finally, a study of the SARS-CoV-2 host interactome was performed, in order to pinpoint basic motifs of its pathogenic course and elucidate its correlation with other interactomes of pathogenic viruses. All these studies led to novel findings of high merit, supporting the efficiency of BioInfoMiner, regarding the reduction of the dimensionality and complexity of various omics datasets. They also confirmed that such a system-level semantic approach may reveal the nodal cellular events and regulatory biomolecules in a phenotypic condition under study and extract critical, novel biological information.

Contents

List of Figures

List of Tables

Chapter 1

Biomedical Ontologies

In the dawn of Big Data in biological sciences, the massive production and analysis of high-throughput data provided new insights about a plethora of biological systems and complex diseases. Applied ontologies and other database systems were developed to accumulate and properly organise the ever-growing body of biological knowledge. The main objective of the present doctoral study is the development of automated computational methodologies for the multi-faceted interpretation and comparative analysis of the results (omics data) of biological high-throughput experiments, exploiting the existing knowledge in widely used biomedical databases. This chapter constitutes an introduction about the origin and the usage of ontologies in biomedical research, delineating the theoretic framework of the whole study.

1.1 The Origin of Applied Ontology

Research paradigms have been evolving since Classical Antiquity, providing various interpretive tools to investigate and understand the nature. After the Scientific Revolution during the early modern period, experimental and theoretical science were the basic research framework to deconstruct and interpret the phenomena [1]. Hundreds of years later, the Digital Revolution furnished the scientific community with powerful computational machines to dive into more intricate systems that were inaccessible before. During the last decades, advances in computational systems in tandem with the avalanche of voluminous digitized data, formed the novel paradigm of data-intensive science [2]. That framework is tightly associated with Artificial Intelligence (AI) and focuses on elucidating

complex scientific inquiries, through the production, process and analysis of large amounts of data. Usually, data-intensive studies exploit the prior knowledge to benchmark or interpret novel data. Prior knowledge is integrated in databases, namely standardized computable structures, easily accessed and manipulated. Applied ontology is one type of these databases, which is used to encode and represent all the available knowledge about a specific scientific discipline in a logic-based format [3].

Seeking the philosophical origins of the term "ontology" one comes across Aristotle's prominent dissertation of Categories [4]. Aristotle proposed the division of *predicamenta* (i.e. the subjects and predicates in a proposition), which correspond to all the ordinary objects of our experience, into ten distinct semantic categories (genera). These categories were divided in sub-classes (species), modeling one of the earliest known semantic classification systems. While Aristotle alleged that genera cannot be joined into a single higher notion, his proposed categorialism seeded the idea of conceptual modeling and hence that of ontology engineering. Despite the latter revisions and rejections from various philosophers, Aristotle's Categories influenced profoundly the evolution of science. His work could be assumed as a progenitor of the applied ontological structures, a major branch in modern sciences, used in various fields.

1.2 Building an Applied Ontology

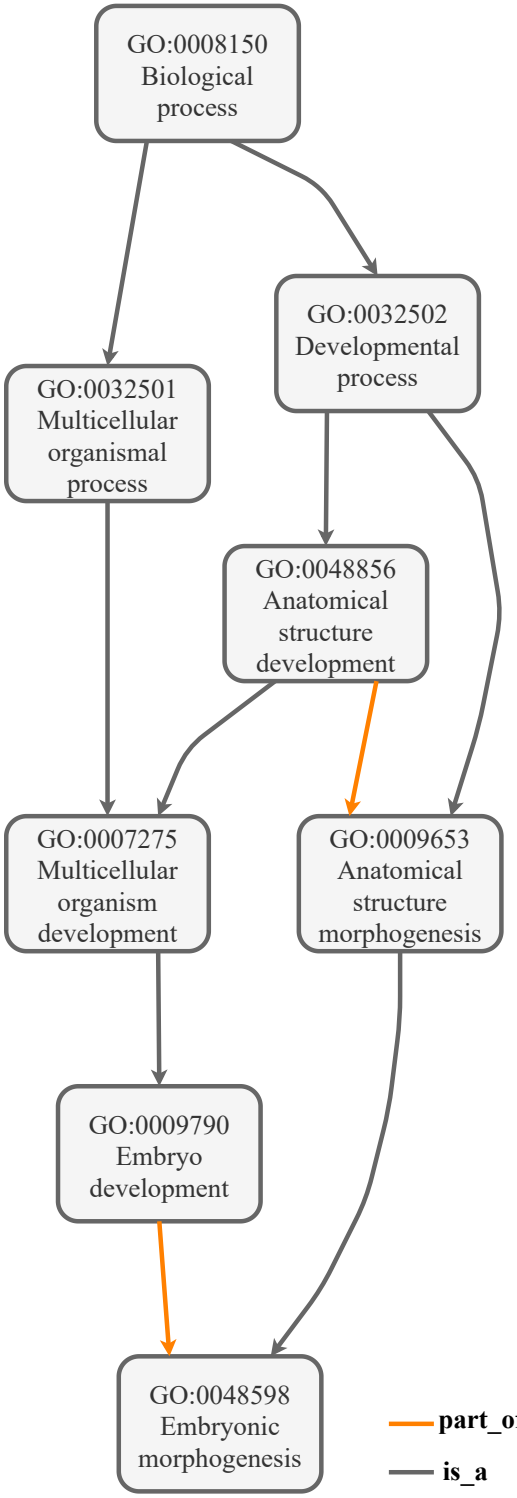
In the field of informatics and AI technologies, ontology is defined as the conceptual approximation of what exists in a certain reality [5]. That approximation is represented by engineering artifacts, targeting to describe effectively the intrinsic nature (types of objects, properties and relations) of the domain of interest (e.g. diseases, molecular reactions, species anatomy), in a machine-readable format. Therefore, the quality of an applied ontology depends on the existing knowledge about the investigated domain and the way that human perception could schematize it. On the other hand, the process of ontology engineering takes into account the functional perspective of an ontology in order to configure its internal structure [6]. A widely used type of applied ontologies is that of controlled vocabulary, which describes the entirety of a concept as a hierarchical collection of lexical terms. The present doctoral study concerns the usage of that type of ontologies in biomedical research. So henceforth, terms like "applied"

and "biomedical" ontologies refer to regimented hierarchical vocabularies, which usually form deep semantic trees to describe various disciplines of biological knowledge.

The principal components of an applied ontology are its classes and their relationships [7]. Each class corresponds to a type of being and a piece of knowledge is a class instance (or object). For example, the Biological Process domain of Gene Ontology [8] describes the universe of biological processes that occur in every living organism. It contains only one class ("biological process") and as for example the phrase "Embryonic morphogenesis" is an instance of that class (Fig. 1-1). In biomedical ontologies, instances are mostly referred to as "terms" and each term has its own identifier and a proper semantic name (label). That name is determined by a non-strict form of Aristotelian *genus-differentia* model [9], combining the names of parental terms and the exclusively differentiated property of that entity. Terms are interconnected with relations, which similarly have their own identifiers. An ontology contains various types of binary relations to precisely manifest the relationships among its terms [10]. Given two terms A and B, their relation is mentioned as "A relation B" and its most frequently used types are the *is_a* (B is the semantic parent of A), *instance_of* (A is an example of class B and not a semantic descendant) and *part_of* (the entity of B encompasses physically that of A). The derived topological space of an applied ontology is a directed acyclic graph (DAG), where nodes correspond to terms and edges to their relations. The most generic ontological genus is located on the graph root and it is gradually explicated in descendant meanings until the leaves, where the most semantically specific terms are arranged.

An applied ontology includes additional features (metadata) in order to completely cover the associated domain and be easily manageable as source of knowledge [11]. The collection of term labels determines its vocabulary, namely the human-readable representation of its terms. Textual descriptions, synonyms, examples, literature references and graphical illustration of ancestral clades, enrich the informational background of terms. Even so, the most significant additional information concerns the association links of terms with entities of other ontologies and databases, facilitating the combination of heterogeneous data and the construction of comprehensive data-intensive frameworks. All the above are formalized and encoded in machine-readable formats, using standardized languages with specific syntax (entities, expressions, axioms), in order to facilitate distribution of ontologies in the community, as well their

integration in complex databases.



Term Metadata

Accession: GO:0048598

Name: embryonic morphogenesis

Ontology: Biological Process

Synonyms: embryonic anatomical, structure morphogenesis

Alternate IDs: GO:0048828

Definition: The process in which anatomical structures are generated and organized during the embryonic phase. The embryonic phase begins with zygote formation. The end of the embryonic phase is organism-specific. For example, it would be at birth for mammals, larval hatching for insects and seed dormancy in plants.

Term Direct Annotation

BMP4: bone morphogenetic protein 4
BMPR1A: bone morphogenetic protein receptor type 1A
C2orf49: chromosome 2 open reading frame 49
CDON: cell adhesion associated, oncogene regulated
FLRT3: fibronectin leucine rich transmembrane protein 3
GLI3: GLI family zinc finger 3
HNF1B: HNF1 homeobox B
HTR2B: 5-hydroxytryptamine receptor 2B
MSX1: msh homeobox 1
MSX2: msh homeobox 2
ROCK1: Rho associated coiled-coil containing protein kinase 1
ROCK2: Rho associated coiled-coil containing protein kinase 2
SHH: sonic hedgehog signaling molecule
SHOX2: short stature homeobox 2
TMED2: transmembrane p24 trafficking protein 2
ZEB1: zinc finger E-box binding homeobox 1
ZEB2: zinc finger E-box binding homeobox 2

Figure 1-1: The tree depicts the ancestral path of ”Embryonic morphogenesis”. ”Biological Process” is the graph root and it is gradually described from top to bottom, whereas in contrast, the ”Embryonic morphogenesis” term encompasses the meanings of its ancestors. Metadata and genomic annotation enrich its informational content.

1.3 Biomedical Ontologies

The massive development of biomedical databases and ontologies is concomitant of the latest advancements in computer science and molecular biology. The role of significant proteins in specific mechanisms had been identified many decades before the epoch of genomics. Their structural and functional characteristics were available only in literature through scientific publications and therefore computationally inaccessible. Initial attempts to organize the biological knowledge in databases started at the end of 20^{th} century. Online Mendelian Inheritance in Man (OMIM) database, which provides a mapping of genetic disorders with their associated human genes, became electronically available in 1987 [12]. The initial version of KEGG, a database for molecular interactions in signaling and metabolic pathways, was published in 1995 [13, 14]. In the same period, EcoCyc database became available, describing the molecular interactions in the metabolic network of Escherichia coli [15, 16]. The genome sequencing of the first eukaryotes, triggered the construction of Gene Ontology (GO), the first ontological vocabulary in biomedical sciences, in 1998 [8, 17]. At the same time, the worldwide network and the progress in computer systems facilitated the wide distribution of biomedical databases among scientists.

However, the completion of Human Genome Project in 2003 [18, 19] marked the dawn of the post-genomic era, disclosing the astonishing complexity of the genomic regulation and human cells functionality [20, 21]. The subsequent paradigm shift in biomedical studies caused a dramatic scientific progress during the next decade, delivering advanced experimental and analytical methodologies to leverage the noted biological complexity and aid the elucidation of intricate biological problems (e.g. complex diseases). In particular, DNA microarray experiments [22, 23] and the recent Next Generation Sequencing (NGS) technologies [24] generated an avalanche of high-throughput experimental genomic data, inaugurating the era of Big Data in biomedicine. Due to the explosion of novel biological data, it was necessary to develop congruous databases to structure and categorize the knowledge, concerning the study of biomolecules in different temporal, spatial and organizational scales. The early success of Gene Ontology [25] inspired the scientific community to embrace and adopt the concept of ontologies in order to describe the novel knowledge in standardized computable structures.

To date, a great repertoire of biomedical ontologies has been created, covering

areas such as cellular functionality [8, 26, 27], diseases [28], phenotypic traits [29–32], organism-specific knowledge [33–36], experimental conditions [37, 38], cell types [39], etc. They became a crucial tool in biological and biomedical research due to their dual benefit. On the one hand, they serve as a mean to structure the knowledge around a specific domain. That perspective spawned the necessity to create rigorous, standardized languages for ontology engineering, such as the Web Ontology Language (OWL) [40] and the Open Biomedical Ontology (OBO) [41], which subsequently led to the construction of well-formed, interoperable ontological schemas. On the other hand, biomedical ontologies are used to annotate genes and gene products, contributing to the holistic description of their role in biological systems. The derived mappings facilitate the semantic reasoning of uncharacterized biomolecules and enable the systemic interpretation of high-throughput omic experiments, as well the integrative analysis over heterogeneous biomedical data. Plenty of algorithms and methodologies have been developed for the exploitation of biomedical ontologies, establishing them as an indispensable part of omics data analysis.

1.3.1 Gene Ontology

Gene Ontology (GO) provides a multi-layered description of gene products' functionality for thousands of species [8]. It came into existence due to the discovery that a large proportion of genes and proteins, which participate in the core processes of eukaryotic cells, are orthologous. Thereby, the functional annotation of a protein in one organism could indicate its profile in other species. Based on that hypothesis, Gene Ontology Consortium was founded in 1998 [17] to construct a unified ontological schema for the annotation of protein functionalities in eukaryotic cells. Thenceforth, the revolution in high-throughput technologies in tandem with the development of advanced computational tools for phylogenetic analysis have enabled the characterization of thousands of species. Nowadays, more than 4500 species have been annotated. GO classes are separated in three main categories: molecular function (MF), biological process (BP) and cellular component (CC). The domain of molecular function refers to the biochemical activities of proteins, while that of biological process indicates the mechanisms where all these functions participate. Cellular Component describes hierarchically the parts of cells where molecular functions occur. The most frequent relations in these sub-ontologies are: *is_a*, *part_of*, *has_part* and

regulates, whereas many terms inherit their meaning from more than one parent (multiple inheritance). Terms metadata contain references to external ontologies and biomedical databases [42–44]. Annotation of proteins is produced through manual curation of experimental results, phylogenetic evidence, authors' and curatorial statements and completely automated computational approaches.

1.3.2 Reactome Database

The Reactome database started as an endeavor to document the molecular interactions in human cells and connect them in the broader network of cellular processes [26]. Nowadays it has been extended to many sequenced species from both eukaryotic and prokaryotic world (e.g. *Mus musculus*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Mycobacterium tuberculosis*, etc). Biomolecules (nucleic acids, proteins, small molecules, complexes), biochemical reactions, pathways and processes constitute its main classes. Different forms of the same physical entity, due to modifications, configurations or activity in various cellular compartments, are accounted as different instances. Reactome follows a bottom-up logic to document the biochemical reactions, while they are chained together by shared physical entities. Reactions comprise small pathways and many of them belong to high-level mechanisms (e.g. glycolysis and gluconeogenesis are part of glucose metabolism, which in turn is a part of metabolism of small molecules). Only the upper levels of that hierarchy (pathways and mechanisms) are associated with *is_a* relations, forming the respective ontological graph.

1.3.3 MGI Mammalian Phenotype Ontology

Mouse has been established as the ideal model organism for the investigation of human biology. For many decades, numerous of research articles provided information about the genetic profiles of diseases, based on experiments on different strains of laboratory mouse. Tons of unstructured and computationally inaccessible data were associating genetic alleles, mutations and environmental factors with abnormal phenotypic traits. At the beginning of 21st century, the consortium of Mouse Genome Informatics (MGI) opted to develop the Mammalian Phenotype (MGIMP) ontology [30], to systematically describe the universe of mouse genetic diseases. Thus, MGIMP ontology started as a standardized vocabulary of abnormal phenotypic traits studied in mouse models. However,

nowadays it has been extended, describing abnormalities appeared to other eukaryotes. All the ontological terms are classified into high-level genera which concern the type of abnormality (morphological, developmental, physiological, behavioral and survival). The underlying graph is constructed with *is_a* relations, while its complexity is increased due to the multiple inheritance. Terms are annotated with the related genotypes, information about the genetic background of studied strains and the experimental conditions, as well with references to other phenotype/disease databases. While the genomic annotation of MGIMP is based on curated results from experiments on laboratory mouse, homology mappings between mouse and other species enable the annotation of a plethora of vertebrates (*Homo sapiens*, *Pan troglodytes*, *Canis familiaris* etc.).

1.3.4 Human Phenotype Ontology

Databases related to human diseases were available online even before the advent of genomics [12]. However, their relational schema was based on simple associations between genetic variations and diseases, whereas redundant annotations and synonymous diseases were impeding any semantic processing. The Human Phenotype Ontology (HPO) project started in 2007 as an effort to create a well-defined ontological schema of individual phenotypic features related to Mendelian (hereditary) diseases and later it was expanded to a broader range of human diseases [29]. Nowadays, HPO constitutes a widespread comprehensive vocabulary used in many clinical and research projects. Its structure is organized as five independent domains that cover different types of phenotype description (frequency, clinical modifier, clinical course, mode of inheritance and phenotypic abnormality), while terms are connected with *is_a* relations. Their annotation includes a wealth of cross-references to other disease-related databases, such as the Unified Medical Language System [45], the Medical Subject Headings [46] and the Disease Ontology [28]. Also, logical definitions based on classes from other ontologies related to anatomy, cell types, cellular functionality, embryology and pathology enable the enrichment of clinical reports. The genomic annotation of terms is mainly produced from manual curation of OMIM [12], Orphanet [47] and DECIPHER [48] databases and text-mining approaches implemented in PubMed.

1.4 Genome Annotation and the True Path Rule

The portrayal of genes and their products is characterised by both quantitative and qualitative properties. Quantitative elements are calculated through experimental techniques and encoded in digitized formats, such as the nucleotide sequence which is represented as a string of letters. However, the functionality or the contribution of a gene in cancer progression cannot be represented in computable format in a straightforward way. That pitfall has been surpassed with the usage of biomedical ontologies, as they compose an suitable semantic framework to define such qualitative features.

The genomic annotation of ontologies could be assigned using either experimental results or reference-based approaches [49]. Experiment-based annotations are generated through the manual correlation of biomolecules with semantic terms, exploiting strong evidences from published studies. In contrast, reference-based approaches propose novel annotations for a biomolecule based on its sequence and structural similarities or phylogenetic relationships with other well-annotated entities. Some methodologies use as reference annotations only those of experiment-based procedures and include various curation steps to provide accurate correlations. Other procedures, which aim to construct genome-scale annotations, are completely automated and despite performing various quality assessments to reduce the erroneous predictions, their output is not manually reviewed. GO adopts all the aforementioned methodologies to provide enriched annotations for the vast majority of sequenced species [50]. Table 1.1 displays the amount of annotations of GO-BP for ten eukaryotic species (version May 2020). Model organisms have greater percentages of experiment-based annotations (e.g. *Arabidopsis thaliana* and *Saccharomyces cerevisiae*) comparing to species of the same taxonomy, which are annotated mainly through phylogeny and automated techniques (*Glycine max*, *Zea mays* and *Aspergillus nidulans* respectively).

A logical consequence of DAG structure is the invention of the True Path Rule (TPR) to correct the genomic annotation [51]. If a term is annotated with a specific biomolecule (positive annotation), then all its ancestral terms must also be annotated with it in a recursive way. Also, if a term is not associated with a biomolecule (negative annotation), then none of its descendant terms could be associated with it (Fig. 1-2). Positive annotations have a bottom-up direction, while negative annotations traverse the tree from top to bottom. That principle enforces the annotation consistency in the whole spread of ontological

tree, as the information flows without gaps. As a consequence, curators focus only on the production of direct annotation and final users need to implement recursively the TPR to correct it (Table Table 1.2).

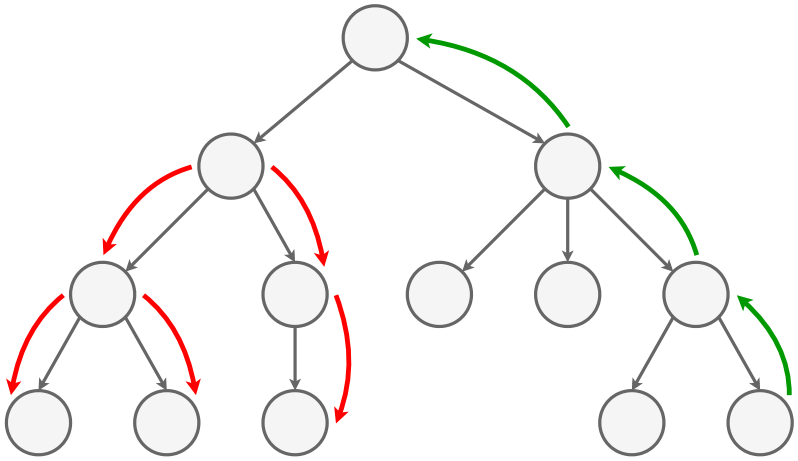


Figure 1-2: Schematic representation of the True Path Rule (TPR). A positive term - biomolecule annotation is adopted by all the ancestral terms (green arrows), whereas negative correlations are universally characterise the descendant branch (red arrows).

Table 1.1: GO-BP Annotations for Different Annotation Procedures

Species	Total	Experimental Evidence	Phylogenetic Evidence	Electronic Evidence
Mus musculus	205779	0.26	0.12	0.40
Homo sapiens	156172	0.28	0.15	0.58
Glycine max	72679	0.00	0.42	0.58
Danio rerio	65738	0.21	0.38	0.42
Zea mays	57976	0.01	0.37	0.62
Arabidopsis thaliana	55328	0.38	0.19	0.43
Oryza sativa	40936	0.03	0.36	0.61
Caenorhabditis elegans	36615	0.33	0.29	0.38
Saccharomyces cerevisiae	33063	0.49	0.20	0.31
Aspergillus nidulans	20814	0.07	0.39	0.53

Table 1.2: True Path Rule Correction In Ontologies Annotation

Species	Ontology	Direct Annotation	TPR Annotation	Rate of Increase
Homo sapiens	GO-BP	135765	1140215	8.4
Homo sapiens	GO-CC	91394	315528	3.5
Homo sapiens	GO-MF	78749	246577	3.1
Homo sapiens	HPO	169281	592650	3.5
Mus musculus	MGIMP	173746	672250	3.9
Mus musculus	Reactome	30330	96927	3.2
Rattus norvegicus	MGIMP	163966	632427	3.9
Bos taurus	Reactome	21232	71230	3.4

1.5 Applications of Biomedical Ontologies

1.5.1 Pathway Analysis

High-throughput, large-scale omic experiments are used for the comprehensive profiling of complex biological systems in different molecular levels (genomic, epigenomic, transcriptomic, proteomic, metagenomic, etc.) [52]. The initial processing of raw data includes specific computational steps depending on the type of omic experiment. Nevertheless, all of them end up to extended lists of genes, transcripts or proteins, with differentiated status among the observed phenotypic categories, or more broadly the examined stratifications. Conventionally, these lists are mentioned as "gene lists", because even if the experiment measures on transcript or protein level, these molecules are usually referred by the name of genes from which they originate. Such a list comprises the first evidence to understand the molecular basis of the underlying phenotypic differentiation. However, further analysis is mandatory to uncover the molecular complexity, untangle its genetic constituents and finally identify or estimate the underlying causal mechanisms.

Pathway analysis is a subsequent computational task, used to look into these questions. It examines the synergistic relations among genes, in order to elucidate the perturbed network of molecular activities and mechanisms, delineating a system-level interpretation [53]. The term "pathway analysis" could be generally used to name any interpretation of a gene list into a set of semantic terms, not

necessarily implying that these terms refer to molecular pathways or functions. In summary, pathway analysis uses the genome annotation of biomedical ontologies to reveal the biological characteristics of the examined phenotype, from a specific semantic perspective. Thereby, it reduces the dimensionality of phenotype description from hundreds or thousands of differentiated biomolecules to tens of semantic terms, such as signaling pathways, molecular functions, phenotypic abnormalities and traits.

A great plethora of pathway analysis tools has been proposed, even though many of them have negligible methodological differences, therefore they could be classified into some particular categories [54]. The final goal is to provide a list of representative terms which have been estimated as significant based on statistical hypothesis tests. These tests evaluate an objective measurement for each term and their null hypothesis could be divided into two main categories: competitive and self-contained [55]. The competitive null hypothesis assumes that the level of differentiation of a gene set, related to a semantic term, is at most as common as that of its complement, namely all the other genes out of that group. The self-contained null hypothesis examines solely the gene set of every term, assuming that it is not differentiated across the given phenotypic classes. All the methods perform simultaneously an independent hypothesis testing for each term. Thus, multiple testing correction (Bonferroni [56], Benjamini-Hochbergand [57]) is usually preferred to eliminate the list of terms which have been falsely promoted as significant (type error I).

Many tools use traditional over-representation statistical tests (hypergeometric, binomial, Fisher's exact and chi-squared) to assess the amount of genes related to a particular term, by calculating the probability to observe that fraction proportionally to its background annotation (competitive null hypothesis) [58–66]. Their input is a predefined list of significant genes (sorted or not), without any additional information about their expression or differentiation status. They are also called "2x2 contingency table methods" as the variables for the underlying statistical tests are derived from the frequency distributions of genes in the input list and background annotation. Due to the TPR (section 1.4), the final lists of over-represented terms could include a lot of redundant meanings, including widespread semantic branches with many parent-child pairs. Some correction methods have been suggested to attenuate the semantic redundancy, exploiting the ontological graphical structure [67, 68]. Other methods adopt the concept of Bayesian inference to construct ad-hoc probabilistic graphical models with genes

and terms as nodes [69–71], in order to extract a representative non-redundant group of significant terms.

Another technique is the gene set analysis (or functional class scoring), whose concept is epitomized by the seminal publication of GSEA [72]. The motivation for the development of these approaches was the thought that weak coordinated changes in functional components might have important role in phenotype physiology [54]. They analyze the list of genes, as it is derived from a high-throughput experiment, without determining the differentiated part based on p-value and fold change. Given the distribution of a gene-level statistic (p-value, fold change, signal-to-noise ratio, z-score, correlation coefficient, coefficient of variation, etc.) a respective pathway-level statistic is calculated for each ontological term (sum, mean, median, maxmean, Kolmogorov-Smirnov, Wilcoxon rank sum, Hotelling's T2, etc.). Then the statistical significance is calculated using permutation tests with either competitive or self-contained hypotheses [72–81].

Topology-based methods constitute the most recent generation of pathway analysis tools. Their algorithmic framework is similar with that of functional class scoring methods, whereas the gene-level statistics are calculated leveraging the topological properties of genes (or their gene products) on the network of pathways or protein-protein interactions (PPIs) [82–89]. Background knowledge-bases do not include only a mapping of terms and gene sets, but also either the topological structure of molecular pathways [14, 26, 90] in specific machine-readable formats (BioPAX [91], SBML [92], KGML [93]) or PPIs data [43, 94, 95]. Gene-level statistics are defined using experimental data in conjunction with gene topological properties, such as centrality measures (degree, betweenness, closeness, etc.), clustering coefficient and shortest path distances. Also some tools construct probabilistic graphical models based on the structure of pathways, to infer those values. Pathway-level statistics are calculated using linear or non-linear aggregation methods or gene set analysis on the node statistics. Then, the statistical significance of terms is calculated using permutation and bootstrap tests, as in the case of gene set analysis.

In the previous years, the laboratory of Metabolic Engineering and Bioinformatics Program (MEBP) of National Hellenic Research Foundation (NHRF; Athens, Greece) proposed the StRAnGER algorithm [96], which belongs to the over-representation methodologies. StRAnGER employs a combination of a parametric (hypergeometric) and a non-parametric statistical test (bootstrap

resampling) [97]. It was developed to perform the pathway analysis utilizing GO and KEGG databases, targeting to reveal the critical aspects of cellular functionality (i.e., nodes higher in GO hierarchy and densely enriched KEGG pathways), rather than semantically specific terms, which are susceptible to the impact of experimental or annotation errors. Initially, the algorithm performs the over-representation test and ranks the terms accordingly. The final ranking is based on a non-parametric, empiric algorithm, which avoids assumptions about the distribution of term enrichments and can be adapted to any ontology. With the bootstrap resampling, StRAnGER avoids the utilization of multiple test correction approaches that are very restrictive and problematic in regard to the finite nature of annotation vocabularies. The main problem of these methodologies is that they tend to promote enrichments yielding very low p-values (close to 0), but holding a very low biological content (e.g. enrichments such as 2/2 and 1/2). The adopted bootstrap procedure provides a corrected measure for the statistical significance of the enrichments based on their frequencies of observation. The algorithm reorders the initial distribution and prioritizes the less frequently observed enrichments. The enrichments are derived as statistically significant if they satisfy both the hypergeometric and the bootstrap thresholds. Hence, terms with less frequent enrichments are prioritized, while they tend to represent broader pathways or functions and, thus, are of stronger biological content.

1.5.2 Gene Prioritization

Yet the derivation of a set of semantic terms through pathway analysis, is only the first step towards a system-level interpretation. The formulation of a concrete network of genes, that captures critical aspects of phenotypic causality, is another important goal. A final list of genes, prioritized and imprinted in a scale free network [98], indicates potential biomarkers and therapeutics targets for diseases and disorders. Gene prioritization approaches focus to provide pertinent targets, leveraging prior knowledge from a variety of biological data sources. Primarily, their reasoning is based on the guilt by association (GBA) principle, namely the assumption that biomolecules interacting in the same neighborhood of a protein network or those which share multiple common properties, it is more likely to impact on the same mechanisms and consequently their modifications are linked with the same phenotypes and diseases. Apart from that, some tools are

based on text-mining algorithms, scanning the scientific literature and disease databases to discover potential gene-disease relations.

In general, a GBA approach prioritizes phenotype candidate genes based on their association with a training gene set, which is assumed as the gold-standard set for the examined phenotype or keywords. The training set could be retrieved from well-curated databases of genetic disorders (OMIM [12], HGMD [99], GAD [100]) or it could be defined by the user. Various methodologies have been proposed to estimate the association of candidate genes and training set under different criteria [101–109]. Generally, they integrate complex, heterogeneous data, such as sequence and protein domain annotation (CDD [110], Pfam [111], InterPro [112]), functional annotation and regulation on molecular pathways (GO [8], Reactome [26], BioCarta [90], KEGG [13, 14], PID [113], PantherDB [114], Pathway Commons [115]), functional coupling networks (FunCoup [116]), transcriptional expression (Atlas [117, 118]), protein-protein interactions and genetic networks (STRING [94], DIP [119], IntAct [43], BioGRID [95]) and sequence homology mappings (HomoloGene [120]). Other methodologies need as input specific keywords or ontological terms instead of seed gene sets [121–127]. These approaches scan the available literature on public repositories (e.g. PubMed [120]) and search for phenotypic annotation in databases (e.g. Ensembl [128], OMIM [12]) to glean potential biomarkers. However, all that techniques are biased by the level of prior knowledge about the examined phenotype, namely the construction of seed genes sets and keyword mappings from previous studies [129].

Another algorithm developed in the MEBP laboratory of NHRF is the GOrevenge [130]. GORevenge focuses to pinpoint and prioritize potential regulatory hub genes, or even further propose potential signatures for the reliable classification of pathological states, in the context of biomarker studies, based on the genomic interpretation of GO graph. Starting from a user-defined set of genes (probably derived from high-throughput experiment or text-mining procedure), GOrevenge delineates their regulatory impact using the annotation of GO domains. Graph-theoretic measures are used to expand the topological range of these sub-graphs and intensify their biological descriptiveness, including more terms from their semantic vicinities. Then the examined gene set is enriched with the genomic annotation of newcomer terms, in order to detect potential hub genes which were not included in the initial set. All genes are sorted according to the amount of their associated GO terms. In an optional

agglomerative process, neighboring terms, considered as functionally similar above a certain threshold, could be incorporated into clusters, eliminating the redundancy of ontological trees. Following that pipeline, GOrevenge reveals specific regulatory genes, which bear significant descriptive information about the considered topology of GO. Comparing to the existing gene prioritization methods, GOrevenge is an unsupervised data-driven approach, which acts as a feature selection filter, ending up to a ranked list of significant genes, without using phenotype related seed genes or keywords.

Chapter 2

Semantic Analysis in Biomedical Ontologies

Essential concepts of semantic analysis on biomedical ontologies are presented in this chapter. Semantic analysis is used in linguistics to reveal relationships between words, sentences and phrases. Similarly, the topological space of biomedical ontologies could be used to explore the relatedness of genes and proteins, based on their semantic representation. The first sections describe the properties of terms on the ontological graph and the idea of shared information between two terms. Then, some well-established pair-wise semantic similarity measures are presented, while the final section referred to the existed approaches to compare sets of terms. The implementation of all the above in functions, as a part of a novel software library, is described in chapter 3.

2.1 Semantic Similarity

The comparison of molecular entities (genes, proteins, RNAs), regardless of the objective criterion, is an essential operation in biomedical studies, applied to detect their degree of conservation, extract motifs, cluster them in families and transfer knowledge from one to another. Methodologies for sequence [131–133] and structure-based comparisons [134, 135] have been developed, even before the era of high-throughput experiments. However, the investigation of their multifaceted biological role was not feasible using these features. The advent of biomedical ontologies assisted to overcome that limitation, as they became the means to profile genes and proteins in various biological domains (molecular

functionalities, diseases, phenotypes, etc.). Additionally, the concept of semantic similarity, borrowed from linguistic studies, provided a straightforward way to conduct quantitative comparisons among biological entities, using their ontological annotations.

In general, semantic similarity indicates the strength of closeness of two meanings, given that they are described in a common knowledge representation system. In biomedical research, such a system is an ontology and its DAG constitutes the respective topological space. Semantic similarity measures take advantage of the topological properties of terms, in order to quantify how close they are. As a consequence, the semantic similarity of two genes or proteins, could be estimated by averaging the pair-wise similarities of their associated terms. In a broader sense, such a comparative analysis could be performed among sets of terms, irrespective of the type of subjects at which these sets are referring to (e.g. proteins, mechanisms, diseases, profile of tumors, etc.). Ergo, semantic analysis is a universal approach, applicable to many different scenarios.

A unified framework has been proposed to explain the reasoning of semantic comparison of two terms [136]. Each term has a unique semantic representation on the graph, while an appropriate function quantifies its specificity (the amount of information). These elements could be considered as the main properties of a term, indicating its unique profile on the DAG. In order to compare two terms, a function is used to define the shared semantic component between them, based on their unique representations. Then, two estimators are used to quantify the specificity of that component (i.e. the amount of shared information) and the degree of divergence. Usually, the semantic similarity measure is as an equation which encompasses these two estimators. As a whole, the semantic properties of terms, the amount of their shared information and the similarity function constitute the core elements of semantic comparison. A plethora of measures have been suggested and their improvement is on an ongoing basis. At this time, there is not any pervasively accepted measure able to overcome in performance all the others in every evaluation study. Also, the ever-increasing comprehension of structural particularities and biases of biomedical ontologies, continuously urge to the development of more sophisticated metrics. As a rule, they are divided into two main categories: edge-based and information-based measures. In this doctorate study only information-based measures were implemented and evaluated.

2.2 Semantic Properties of Terms

Ontological terms are characterized by information-based and graph-based attributes. Their combination determines a semantic entirety on the graph. Each term has a set of parents (ascendant terms directly related with it), an ancestral branch (ancestors; all the passed terms traveling from its vertex to the root), a set of children (directly related descendant terms; for leaves this set is empty) and its descendants (the offspring graphical branch) [137]. Another important property is their genome annotation which is described in section 1.4. All the above delineate the semantic profile of a term and provide the necessary information to determine some additional quantitative features.

The Information Content (IC; Shannon Information) measures the specificity of a term [138]. Intuitively, it is inversely related to the probability of occurrence on the graph. That probability is calculated either based on the existence of the meaning of term in the ontology (which is mentioned by the amount of its descendants; internal, graph-based approach) or the amount of the associated genes in the annotation (external, annotation-based approach). The IC of a term t is defined as the negative log likelihood of probability of its occurrence:

$$IC(t) = -log_2\left(\frac{|N_t| + 1}{|N_{total}|}\right) \tag{2.1}$$

According to the graph-based approach N_t is the set of descendants and N_{total} is the complete set of ontological terms. In annotation-based approach N_t indicates the associated genes and N_{total} is the whole set of annotated genes. By definition, $IC(t)$ is monotonically increasing from root to the leaves.

However, IC is not able to point out the topology of a term on DAG, as it takes into consideration only its degree of occurrence. Thereby, leaf nodes have maximum information content concerning the graph-based method, albeit they are located in different depths, depending on the quality of annotation and the curation in distinct graph areas. Another measure was proposed to surpass that limitation. Initially, Song et al. [139] proposed the concept of Semantic Weight (SW), which is inversely correlated with the IC of a term t:

$$SW(t) = \frac{1}{1 + exp\big[-IC(t)^{-1}\big]} \tag{2.2}$$

Then, they defined the Semantic value (SV) of term t, an aggregated semantic

contribution of all its ancestors, which depicts its semantic distance from the root node:

$$SV(t) = \sum_{a \in Anc(t)} SW(a) \tag{2.3}$$

The function of $SV(t)$ is linearly correlated to the cardinality of ancestors set. Extremely great scores imply extensively described regions, where terms have plenty of ancestors comparing to other more succinct branches. An important difference between IC and SV is that SV is always greater than zero, instead of IC, whose logarithmic function zeroes the value of root node. Taking into account the assumption that every node is both ancestor and descendant of itself, none of the nodes has empty set of ancestors, so SV ranges in the set of positive real numbers.

Studying the structure of biomedical ontologies, Mazandu and Mulder [140] mentioned that there are terms which are divided into only one descendant meaning, without spreading sideways. Such a term should be topologically identical with its exclusive child. Both $IC(t)$ and $SV(t)$ functions lack to discern such peculiarities, scoring higher specificity value for child term. Topological Position (TP) was invented as an alternative measure to reflect the correct topological location of a term:

$$TP(t) = \begin{cases} 1, & \text{if t in the root} \\ \prod_{a \in Anc(t)} \frac{TP(a)}{|a_{children}|}, & \text{otherwise} \end{cases} \tag{2.4}$$

It is evident that $TP(t)$ is monotonically decreasing in top-down direction. For a parent term, that property is divided uniformly according to the number of its children and thus the more children a term has, the greater their topological difference will be. In addition, if two terms have the same parents then they are topologically identical. TP could be used to replace the probability of occurrence in the logarithmic function of IC, providing a specificity measure strictly based on topology (topological information).

2.3 Shared Information of Two Terms

Conceptually, the semantic representation of a term could be determined by its ancestors. Hence, the shared semantic component of two terms is related to the common part of their ancestral paths. Working on the semantic analysis in WordNet, Resnik [138] proposed that the Shared Information (SI) of two words, and consequently that of two ontological terms t_1 and t_2, could be estimated by their most informative common ancestor (MICA), namely their common ancestor with the highest IC value:

$$SI_{MICA}(t_1, t_2) = IC\big[MICA(t_1, t_2)\big] \tag{2.5}$$

The above assumption was adopted by many semantic similarity measures. However, it discards the semantic contribution of other common ancestors, as many terms in biomedical ontologies inherit their meaning from disjunctive ancestral paths. That weakness led to the construction of more sophisticated methods to detect the shared semantics. GraSM [141] was the first approach which adopted the concept of disjunctive ancestors (DA). It suggests that two ancestors a_1 and a_2 of term t are disjunctive if there is a path from a_1 to t not containing a_2 and a path from a_2 to t not containing a_1:

$$DA_{GraSM}(t) = \Big\{\big(a_1, a_2\big) | \big[relation(a_1, t) \wedge relation(a_2, t)\big]\Big\} \tag{2.6}$$

where $relation(a_1, t) = \big[\exists p : \big(p \in Paths(t, a_1)\big) \wedge (a_2 \notin p)\big]$ and likewise for $relation(a_2, t)$. Given two terms t_1 and t_2 and their common ancestors $CA(t_1, t_2)$, their disjunctive common ancestors (DCA) are defined as the set of the most informative terms which are included in the union of their disjunctive ancestors:

$$DCA_{GraSM}(t_1, t_2) = \Big\{a_1 | a_1 \in CA(t_1, t_2) \wedge \forall a_2 : [relation_A \Longrightarrow relation_B]\Big\} \tag{2.7}$$

where $relation_A$ states that a_2 is different from a_1 and it has equal or greater IC value than a_1: $relation_A = \big[\big(a_2 \in CA(t_1, t_2)\big) \wedge \big(IC(a_1) \leq IC(a_2)\big) \wedge \big(a_2 \neq a_1\big)\big]$, while $relation_B$ implies that a_1 and a_2 are disjunctive ancestors for either t_1 or t_2: $relation_B = \big[(a_1, a_2) \in (DA_{GraSM}(a_1) \cup DA_{GraSM}(a_2))\big]$. Thus, examining the set of common ancestors in a descending order, according to their IC values, an ancestor will be assigned as disjunctive if and only if there is a path from

one of the two terms to that ancestor, distinct of any other path from that term to the already detected disjunctive ancestors. The respective SI is equal to the average IC of $DCA_{GraSM}(t_1, t_2)$:

$$SI_{GraSM}(t_1, t_2) = \frac{1}{|DCA_{GraSM}(t_1, t_2)|} \sum_{a \in DCA_{GraSM}} IC(a) \tag{2.8}$$

The GraSM algorithm has computational complexity $O(n^3)$, which strongly limits its implementation in large-scale studies. Furthermore, that approach penalizes the existence of multiple inheritance, as it decreases the amount of SI comparing to the MICA approach. That behavior is proper when one of the terms has multiple parents, which increase its semantic divergence from the other term. However, it disagrees with the human perception when the disjunctive common ancestors are inherited in parallel. Targeting to overcome that drawback, Couto and Silva proposed the Disjunctive Shared Information (DiShIn) algorithm [142] to estimate the SI of two terms, using as criterion the number of distinct paths between these terms and their common ancestors. For each ancestor, DiShIn calculates the amount of paths that connect it with the investigated terms and then defines their absolute difference: $PD(a) = |Paths(t_1, a) - Paths(t_2, a)|$. An ancestor is assumed as disjunctive if it is the most informative one with that value of paths difference:

$$DCA_{DiShIn}(t_1, t_2) = \Big\{ a_1 | a_1 \in CA(t_1, t_2) \wedge \forall a_2 [relation_A \Longrightarrow relation_B] \Big\} \tag{2.9}$$

where $relation_A$ is true for other common ancestors with the same PD value: $relation_A = \big[\big(a_2 \in CA(t_1, t_2)\big) \wedge \big(PD(a_1) = PD(a_2)\big) \big]$ and $relation_B$ assures that a_1 is the most informative with that PD value: $relation_B = [IC(a_1) > IC(a_2)]$. While the computational complexity is still $O(n^3)$, the calculation of PD values could be performed once, when the ontological graph is updated. Seemingly, DiShIn does not penalize the existence of common multiple inheritance. However, it is sensitive to complex ontological graphs, which include a lot of vertices with multiple inheritance and subsequently the amount of paths that start from a term increases eminently moving to the upper tiers of graph. That behavior might cause unique PD values for generic terms and falsely promote them as

disjunctive ancestors. The respective SI could be quantified similarly to GraSM:

$$SI_{DiShIn}(t_1, t_2) = \frac{1}{|DCA_{DiShIn}(t_1, t_2)|} \sum_{a \in DCA_{DiShIn}} IC(a) \tag{2.10}$$

Finally, Mazandu and Mulder [143] alleged that all the informative common ancestors should be considered, instead of defining the DCA set. This approach is referred as eXtended GraSM (XGraSM) and its computational complexity is constant, as it simply aggregates the weighted IC values of all the common ancestors:

$$SI_{XGraSM}(t_1, t_2) = \frac{1}{|CA(t_1, t_2)|} \left[1 + \sum_{a_j \in CA(t_1,t_2) \wedge a_i \neq MICA}^{n-1} \frac{IC(a_j)}{IC(MICA)} \right] \tag{2.11}$$

2.4 Semantic Similarity Measures for Ontological Terms

As it is mentioned in section 2.1, semantic analysis could be performed using edge-based or information-based methodologies. Mainly, edge-based measures define the semantic distance between two terms as the number of edges in the path between them, or if there are many paths, by the average, maximum or minimum value. Semantic similarity is simply the inverse score of their semantic distance. More sophisticated approaches use different edge weights to reflect some degree of graphical depth and assign higher similarities near leaves, comparing to the upper tiers of graph, for the same amount of edges. However these measures adopt two improper assumptions about the structure of biomedical ontologies [144]: uniformly distributed nodes and semantically equal edges in a specific depth of the whole tree. These limitations had caused the development and dominance of information-theoretic approaches.

Information-theoretic measures assume that the amount of SI of two terms is an estimator of their semantic similarity. Primarily, all the proposed measures adopted the concept of MICA to quantify the SI of two terms. In general, any method that quantifies the SI of two terms could be used. Resnik [138] was the first who proposed such a measure, defining the semantic similarity of two terms

equal to the IC of their MICA:

$$SST_{Resnik}(t_1, t_2) = IC\big[MICA(t_1, t_2)\big] \tag{2.12}$$

The principal limitation of Resnik's measure is that it ignores the specificity of the compared terms [145]. As a result, a pair of generic terms has the same semantic similarity with a pair of extremely specific terms, if they have the same shared semantic component (MICA or DCA). However, this is not consistent with the human perception [139], which suggests that the semantic similarity of generic terms should be greater, as they are semantically broader entities. Other approaches were proposed to normalize Resnik's equation exploiting the specificity of terms:

$$SST_{Lin}(t_1, t_2) = \frac{2 \times IC\big(MICA(t_1, t_2)\big)}{IC(t_1) + IC(t_2)} \tag{2.13}$$

$$SST_{JiangConrath}(t_1, t_2) = 1 - \Big[IC(t_1) + IC(t_2) - 2 \times IC\big(MICA(t_1, t_2)\big)\Big] \tag{2.14}$$

$$SST_{Schlicker}(t_1, t_2) = \frac{2 \times IC\big(MICA(t_1, t_2)\big)}{IC(t_1) + IC(t_2)} \times [1 - P_{occurrence}\big(MICA(t_1, t_2)\big)] \tag{2.15}$$

$$SST_{Faith}(t_1, t_2) = \frac{IC\big(MICA(t_1, t_2)\big)}{IC(t_1) + IC(t_2) - IC\big(MICA(t_1, t_2)\big)} \tag{2.16}$$

$$SST_{Nunivers}(t_1, t_2) = \frac{IC\big(MICA(t_1, t_2)\big)}{max\big[IC(t_1), IC(t_2)\big]} \tag{2.17}$$

A deeper insight into the above measures unveils that they are analogous to some widely used measures of Set Theory [146]. Resnik's similarity is tantamount with the cardinality of intersection of the two compared sets. Likewise to Resnik's limitation, intersection set produces biased similarity estimations and a correction with the union or the sizes of individual sets is necessary. Lin's measure [147] uses the Sørensen–Dice coefficient formula, while Schlicker [148] proposed a weighted version of it, using the probability of occurrence of MICA. Jiang and Conrath [149] defined the semantic distance (the second factor of equation 2.14) as the amount of mutually exclusive information, which reminds the Manhattan

(Hamming) distance of two sets. Faith [150] measure is equivalent to the Jaccard index, also known as "intersection over union" score. Nunivers [143] uses the equation of Braun-Blanquet coefficient, where the IC of MICA is normalized by the maximum IC value of the compared terms.

Another measure similar to the Sørensen–Dice coefficient was proposed by Wang et al [151] and later it was revised by Song et al [139] (Aggregate Information Content - AIC). It uses the concept of SV (equation 2.3), in order to overcome the limitations of IC. AIC defines the SI (nominator) as the sum of the SW values (equation 2.2) of common ancestors, while the denominator is the aggregated semantic contribution of all ancestors:

$$SST_{AIC}(t_1, t_2) = \frac{\sum_{a \in Anc(t_1) \cap Anc(t_2)} 2 \times SW(t)}{SV(t_1) + SV(t_2)} \tag{2.18}$$

2.5 Semantic Similarity Measures for Gene Products

The idea of semantic comparison of ontological terms is expanded in groups of terms, which could constitute the semantic description of biomolecules, phenotypes, diseases, etc. In such a way, the semantic similarity of two gene products could be estimated by comparing their associated term sets. These measures are separated into two categories: pair-wise and group-wise [145]. Pair-wise measures combine the semantic similarities of terms in order to extract an averaged value for the compared concepts. On the other hand, group-wise approaches do not use the pair-wise similarities but compare the sets of objects and estimate their likeliness with traditional binary similarity measures.

Given two gene products g_1 and g_2 and their direct annotations (T_1, T_2) with cardinalities N and M respectively, the average and maximum semantic similarity scores are defined as [152]:

$$SSG_{AVG}(g_1, g_2) = \frac{1}{N \times M} \sum_{t_i \in T(g_1), t_j \in T(g_2)} SS(t_i, t_j) \tag{2.19}$$

$$SSG_{MAX}(g_1, g_2) = max\big\{SST(t_i, t_j) : t_i \in T(g_1) \wedge t_j \in T(g_2)\big\} \tag{2.20}$$

The above measures are sensitive to outliers and multi-modal similarities distri-

bution and consequently they are more susceptible to output biased or completely erroneous results. To surpass that limitation, other measures construct two vectors of maximum similarity scores, one for each set of terms T_x. Specifically, each term $t_x \in T_x$ is compared with the terms in set T_y and the maximum similarity is returned: $S\big(t_x, T_y\big) = max\Big\{SST(t_x, t_y) : t_y \in T_y\Big\}$. Various normalization techniques have been proposed to average these two vectors and derive the semantic similarity of gene products. The Best Match Average (BMA) [153] approach calculates the mean maximum score for each set of terms and then averages the two means:

$$SSG_{BMA}(g_1, g_2) = \frac{1}{2}\Bigg(\frac{1}{N}\sum_{t_i \in T(g_1)} S\big(t_i, T(g_2)\big) + \frac{1}{M}\sum_{t_j \in T(g_2)} S\big(t_j, T(g_1)\big)\Bigg) \tag{2.21}$$

The Best Match Maximum (BMM) (also known as RCMax [154]) returns the maximum mean score, instead of their average value:

$$SSG_{BMM}(g_1, g_2) = max\Bigg\{\frac{1}{N}\sum_{t_i \in T(g_1)} S\big(t_i, T(g_2)\big), \frac{1}{M}\sum_{t_j \in T(g_2)} S\big(t_j, T(g_1)\big)\Bigg\} \tag{2.22}$$

Finally, the Average Best Matches (ABM) is the mean of best matches [153]. The sum of maximum scores is divided by the sum of set sizes:

$$SSG_{ABM}(g_1, g_2) = \frac{1}{N+M}\Bigg(\sum_{t_i \in T(g_1)} S\big(t_i, T(g_2)\big) + \sum_{t_j \in T(g_2)} S\big(t_j, T(g_1)\big)\Bigg) \tag{2.23}$$

Group-wise measures exploit the annotation of a gene product as a set of objects and not as a semantic representation. Thus they use the whole part of ontological graph which is annotated with a gene product, instead of its direct annotation. Like the measures for the comparison of term pairs, group-wise approaches use formulas of Set Theory (Sørensen–Dice, Jaccard and Braun-Blanquet coefficients). Terms in the compared sets ($A(g_1)$, $A(g_2)$) could be represented as equivalent or weighted entities (based on their IC) [152]:

Sørensen–Dice formula:

$$SSG_{DIC}(g_1, g_2) = \frac{2 \times \sum_{t_s \in A(g_1) \cap A(g_2)} IC(t_s)}{\sum_{t_i \in A(g_1)} IC(t_i) + \sum_{t_j \in A(g_2)} IC(t_j)} \tag{2.24}$$

$$SSG_{DB}(g_1, g_2) = \frac{2 \times \left|A(g_1) \cap A(g_2)\right|}{\left|A(g_1)\right| + \left|A(g_2)\right|} \tag{2.25}$$

Jaccard-index formula:

$$SSG_{GIC}(g_1, g_2) = \frac{\sum_{t_s \in A(g_1) \cap A(g_2)} IC(t_s)}{\sum_{t_u \in A(g_1) \cup A(g_2)} IC(t_u)} \tag{2.26}$$

$$SSG_{UI}(g_1, g_2) = \frac{\left|A(g_1) \cap A(g_2)\right|}{\left|A(g_1) \cup A(g_2)\right|} \tag{2.27}$$

Braun-Blanquet formula:

$$SSG_{UIC}(g_1, g_2) = \frac{\sum_{t_s \in A(g_1) \cap A(g_2)} IC(t_s)}{max\left\{ \sum_{t_i \in A(g_1)} IC(t_i), \sum_{t_j \in A(g_2)} IC(t_j) \right\}} \tag{2.28}$$

$$SSG_{UB}(g_1, g_2) = \frac{\left|A(g_1) \cap A(g_2)\right|}{max\{\left|A(g_1)\right|, \left|A(g_2)\right|\}} \tag{2.29}$$

Chapter 3

BioInfoMiner: Interpretation of Cellular Complexity through Semantic Network Analysis

This chapter presents the software architecture of a novel automated workflow for the semantic interpretation of omics data, named BioInfoMiner, whose initial version was published in [155]. It combines computational methodologies and data visualization techniques, in order to deliver comprehensible illustrations of cellular complexity. The workflow executes sequentially pathway analysis and gene prioritization. Starting from a list of individual genes, associates them with the relevant semantic terms and manages to prioritize them according to their involvement in the cellular phenotype under study. BioInfoMiner is designed to use various ontological vocabularies for different organisms. The final results are visualized in an interactive bipartite graph, embedded in a modern front-end interface.

3.1 Introduction

As it is mentioned in section 1.5, the study of complex phenotypes with modern high-throughput experiments requires the systemic interpretation of results, in order to pinpoint and understand the synergies of molecular events that give rise to the crucial phenotypic traits. Various pathway analysis methodologies have been proposed to provide related solutions. Additionally, gene prioritization techniques focus on clarifying the most important molecular factors, given an

ad-hoc reference condition. Targeting to provide a unified solution for the above tasks, BioInfoMiner, a novel automated workflow was constructed, based upon the StRAnGER and GOrevenge algorithms, presented in section 1.5.

The initial version of StRAnGER performs pathway analysis on human and mouse omics data, using GO and KEGG annotations. The core module was programmed in Perl, while MATLAB (R2007a) Bioinformatics Toolbox was used to handle the graphical data of results. GOrevenge was developed to exploit the GO annotation of an input gene list and detect genes which have preeminent centrality on the underlying topology of GO graph. While its concept is generic and applicable to any biomedical ontology, GOrevenge has been designed to analyze only human and mouse data with GO annotation. Its version was developed with the Python language and it was integrated in a web-application architecture.

Intuitively, the list of over-represented terms, defined by StRAnGER, reflects the system-level functional entities (GO and KEGG terms), involved in the manifestation of the given phenotype. The individual genes are linked into broader entities (mechanism and functions), which shape a unique semantic description. In addition, GOrevenge serves as a prioritization tool to unveil genes which bear significant descriptive information and are linked with semantically distant entities on the semantic topology. Setting these approaches in a sequential mode (Fig. Fig. 3-1), their implementation ends up to a unique semantic graph, connected with a signature of genes. The comprehensive integration of these results could be realized as a bipartite graph, which illustrates the over-represented terms on the DAG and their hidden associations with the prioritized genes. The objective in [155] was to design an appropriate software library in order to construct such a workflow, facilitating the integration of novel ontologies and annotations for all the available species and providing it as an unsupervised and automated application.

3.2 Methods

Apart from the construction of the workflow of BioInfoMiner, that work targeted to create a flexible framework, able to integrate and use different ontologies and annotations in interoperable data formats, without the need to configure the back-end procedures in each case. Taking into account the above, it was necessary to

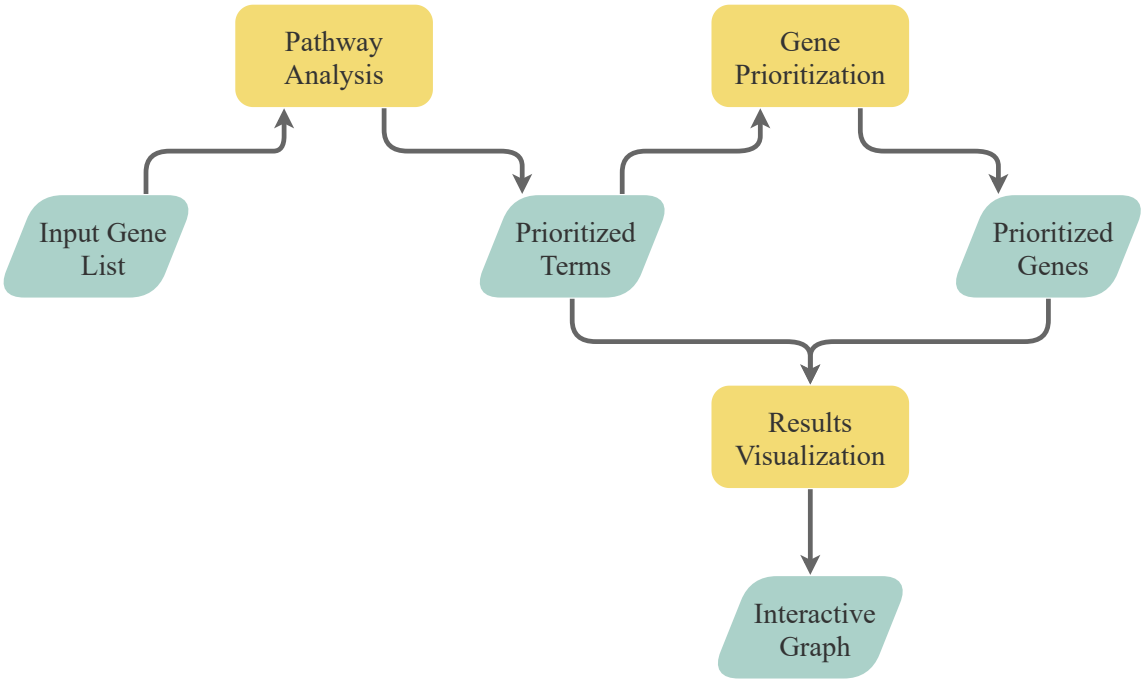


Figure 3-1: BioInfoMiner workflow. The sequential implementation of StRAnGER and GOrevenge algorithms ends up to a prioritized set of over-represented semantic terms and its relevant list of linker genes. Results are presented as a bipartite interactive graph.

create a modular software library to facilitate the reuse of various modules and data types in different processes and for different purposes. Consequently, the library was designed in object-oriented logic, using Python 2.7. The following sections describe the novel features of each algorithmic step and the most important Python classes of software library, used both in computational analysis and data visualization.

3.2.1 Modifications of the Existed Methods

StRAnGER and GOrevenge algorithms were designed from the ground up to build their internal computational processes in a common object-oriented framework. In the updated version of StRAnGER, the algorithmic process did not change substantially, however the new implementation facilitates the integration of novel ontologies and annotations. Also, the statistical tests were parallelized, using the multiprocessing Python library. The crucial step of GOrevenge is the detection of genes, which are highlighted as hub nodes on the considered semantic graph. Its previous version was executing two agglomeration procedures (Graph-BubbleGO and Resnik-BubbleGO) to expand the plexus of that graph

with neighbor terms, in order to avoid potential pitfalls and gaps in the genomic annotation. The new version uses the corrected genomic annotation, derived from the adoption of the TPR (section 1.4), so that step was useless and it was removed. Gene prioritization was being implemented either on the naive graph or pruned versions of it, where the amount of terms was being reduced in arbitrarily defined proportions: 90%, 60% and 30%. In this way, the final output was consisting of four different prioritized lists, letting the user to decide the most preferable. Moreover the algorithm was configured to take into account the total background annotation of the considered graph and not the input list of genes. Thus, common master regulatory genes were more likely to be inclined as hub genes. That step was completely changed, as the purpose of the novel workflow was to provide a unique, unbiased signature, including genes solely from the initial input list. These genes are associated with the quasi-uncorrelated components of the ontological graph, consisting a specified signature for the examined case. For that reason, a novel clustering approach was developed to reduce graph dimensionality into an optimal set of distantly-related ontological components, using semantic analysis.

3.2.2 The GraphNode Class

Each ontological term is represented as a *GraphNode* object. Its attributes contain specific semantic properties (parents, ancestors, children, descendants, types of relations, metadata). The number and the length of paths that pass through the specified term and its ancestors are stored in appropriate data types. While all that information is captured with non-linear complexity recursive algorithms, *GraphNode* instances are created once, when the ontology is updated, during the construction of the respective *Graph* object.

3.2.3 The Graph Class

The *Graph* class is used to imitate the structure of an ontological graph (DAG) for a certain organism. Thereby, the *Graph* instance of GO-BP for *Homo sapiens* is different from that of *Mus musculus*. An instance is constructed once, parsing the OBO format file of the respective ontology, using custom Python modules. The genomic annotation of species is retrieved from the Ensembl database [128], using the biomaRt package of R/Bioconductor [156]. A *Graph* object includes

all the annotated terms as *GraphNode* instances. Its methods are used for many internal tasks, such as the correction of genomic annotation through the implementation of the TPR. Each *Graph* object is updated when there is a new version of ontology structure or genome annotation. Then, it is transformed in binary format, using the cPickle serialization module of Python, and it is stored as a collection in MongoDB database.

3.2.4 The Semantics Class

Semantics is a complex class, which includes all the necessary methods to perform semantic analysis on the ontological graphs. It operates as a wrapper module to perform various semantic operations using the attributes of a *Graph* object. Its methods are inherited from multiple smaller classes (Fig. Fig. 3-2). Specifically, it inherits methods and attributes from two types of classes: Base and Operator. Each Base class contains a list of methods for a specific task of semantic analysis (e.g. calculation of similarity between two terms). One or more Base classes are incorporated in an Operator class, which includes high-level methods to execute the considered tasks given the appropriate parameterization (e.g. given the label of a pair-wise measure as input, it calls the proper inherited method to calculate the similarity of two terms). Using that model, each Base class could be developed and improved separately, without changing or overwriting data types of other classes. Their methods are always called indirectly using the wrappers of Operator class.

Methods of *SemanticsBase* and *SemanticsAncestorsBase* are called through *SemanticsPropertiesOperator*. That module is used to retrieve the semantic properties of a term (section 2.2) and to define the shared semantic component of two terms, based on their ancestral paths (section 2.3). As it is mentioned in section 2.4, many formulas of binary measures of Set Theory are used by information-based semantic similarity measures. *BinaryMetricsBase* contains these formulas and *SemanticsPairwiseMetricsBase* is used to call them and define the corresponding semantic similarity measures. On the other hand, *AggregateMetricsBase* contains functions to calculate the semantic similarity of gene products (section 2.5). The intermediate "Operator" classes are combined into the advanced *Semantics* class which inherits all the methodologies to perform various semantic analysis tasks, given an input *Graph* object.

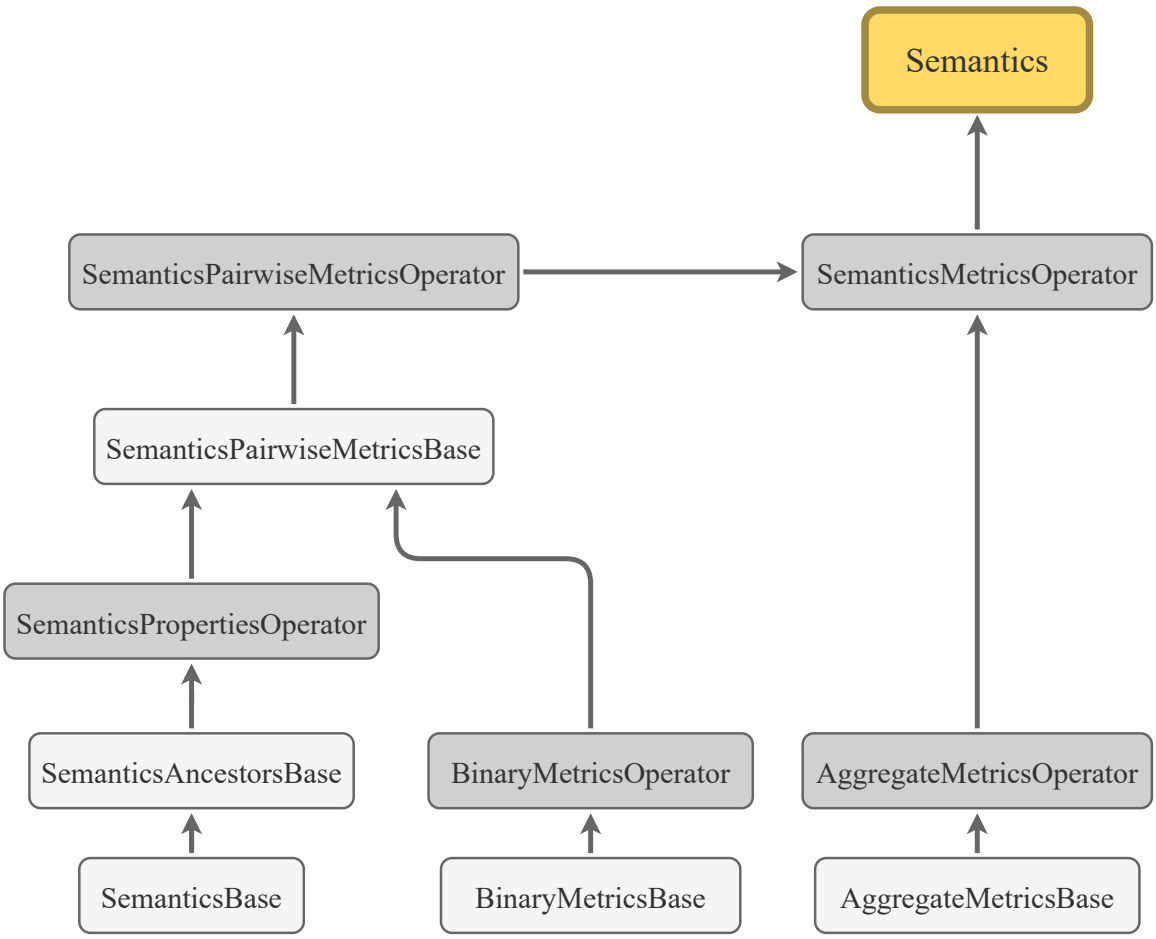


Figure 3-2: The architecture of the *Semantics* class. Using multiple inheritance, various Base and Operator classes are combined to construct a comprehensive powerful class, able to perform semantic analysis on an input *Graph* object.

3.2.5 Classes for Semantic Clustering

GraphClusteringOperators class contains various algorithms for the semantic clustering of terms. Their input is a *Graph* object, which is used to define an internal instance of the *Semantics* class. Different types of clustering could be performed (agglomerative, network-based) and the final output consists of a group of clusters (*GraphCluster* instances). Terms clustered together could be retrieved from the attributes of the respective *GraphCluster* object.

3.2.6 The Computational Workflow

As described above, the computational workflow executes sequentially pathway analysis and gene prioritization. A specific module was constructed for each task to include all the necessary methods (Fig. A-1 A, Fig. A-1 B). *PathwayAnalysisModule* loads and deserializes the *Graph* object from MongoDB, regarding the user-defined parameters (ontology and organism). The TPR-corrected genomic annotation of *Graph* is used to perform the analysis. The extracted list of statistically significant enriched terms serves as input to *GenePrioritizationModule* for the execution of gene prioritization. A semantic clustering of terms is implemented, using one of the *GraphClusteringOperators* classes. The output consists of a set of distinct ontological clusters (structured as *GraphCluster* instances) and the associated hub genes (the gene signature), prioritized according to their involvement in the underlying topology (Fig. A-1 C).

3.2.7 Visualization of Results

In order to illustrate the computational analysis results, in combination with experimental data, a visualization module was developed (Fig. 3-3). The main idea was to illustrate in a common framework the over-represented part of the ontological tree (Fig. 3-4 A), its segmentation into the distinct semantic clusters (Fig. 3-4 B) and their associations with the prioritized genes. Such a comprehensive plot could be represented as a hierarchical graph, where nodes (i.e the enriched terms) are clustered on broader entities (not necessarily containing nodes of the same branch) and each entity is linked with external nodes (genes), shaping an intuitive bipartite graph. Also, the visualization in an interactive front-end environment requires the transformation of data into JSON objects.

In order to carry out the above idea, a novel, automated workflow for data visualization was developed. Initially, a *Graph* object is generated and the over-represented terms in combination with their ancestors are marked to be visualized. Usually, the amount of these terms and the degree of their interconnections, produce voluminous trees, making the neat representation of graph structure practically elusive. Thus, a preliminary pruning is performed, clustering all the over-represented terms using methods of *GraphClusteringOperators*. Each pair of terms is connected only with its MICA, in an iterative process which terminates to the graph root. The output hierarchy represents a succinct, non-redundant version of the over-represented graph, as each term has a single parent and

intermediate semantic branches with non significant terms are condensed into single nodes.

In the second step, the Graphviz software [157] is used to create the visual representation of the extracted tree. Graphviz contains open-source tools to represent graphs and networks in simple two-dimensional plots, using the DOT scripting language, while its methodologies are available as Python libraries (graphviz package). The output SVG (Scalable Vector Graphics) object contains the coordinates of graph nodes and edges. A Python XML parser (xml.dom.minidom) is used to read the SVG and extract these data in the appropriate format.

During the step of gene prioritization, the over-represented terms are classified into distinct semantic clusters. As a result, each term and subsequently its point in the new two-dimensional topology belongs to a broader entity. The underlying region of each cluster is demarcated by applying the Voronoi decomposition algorithm [158] via the spatial.Voronoi module of SciPy Python library. In general, the Voronoi algorithm segregates a plane into partitions, according to a given set of objects (seed points). In this implementation, it is used to determine the individual polygons of each semantic cluster (Fig. 3-4 B). Finally, the coordinates of graph elements and polygons, as well the associations of prioritized genes with terms and clusters and other useful data are organised in JSON format objects.

The visual representation is constructed using the JavaScript language and the D3.js library, which is suitable to visualize data as SVG elements and assigning specific functionalities to them. Each visualized entity (i.e. graph nodes, edges, semantic clusters, prioritized genes) is a unique SVG object, having its own attributes and methods, so it could be manipulated separately. Also each one contains internally a bulk of data which enrich its informational content and define hidden associations with other visualized entities. Each node, in the constructed hierarchical tree, constitutes a distinct circle. If the respective term is noted as over-represented, then the node will be a D3 pack layout, including the associated genes as internal circles Fig. A-2. The user is able to navigate and disclose the genomic content of each term using zoom-in and zoom-out functionalities. The polygons of each semantic cluster are uniquely colored. The user is able to focus on the content of specific cluster using zoom methods (Fig. A-3 and A-4). Below the area of the hierarchical tree, there is another SVG object, which displays the prioritized genes as circles. The mouse-over functionality reveals through highlighting their associations with the semantic

clusters and subsequently their impact on the underlying semantic topology (Fig. A-5 and A-6).

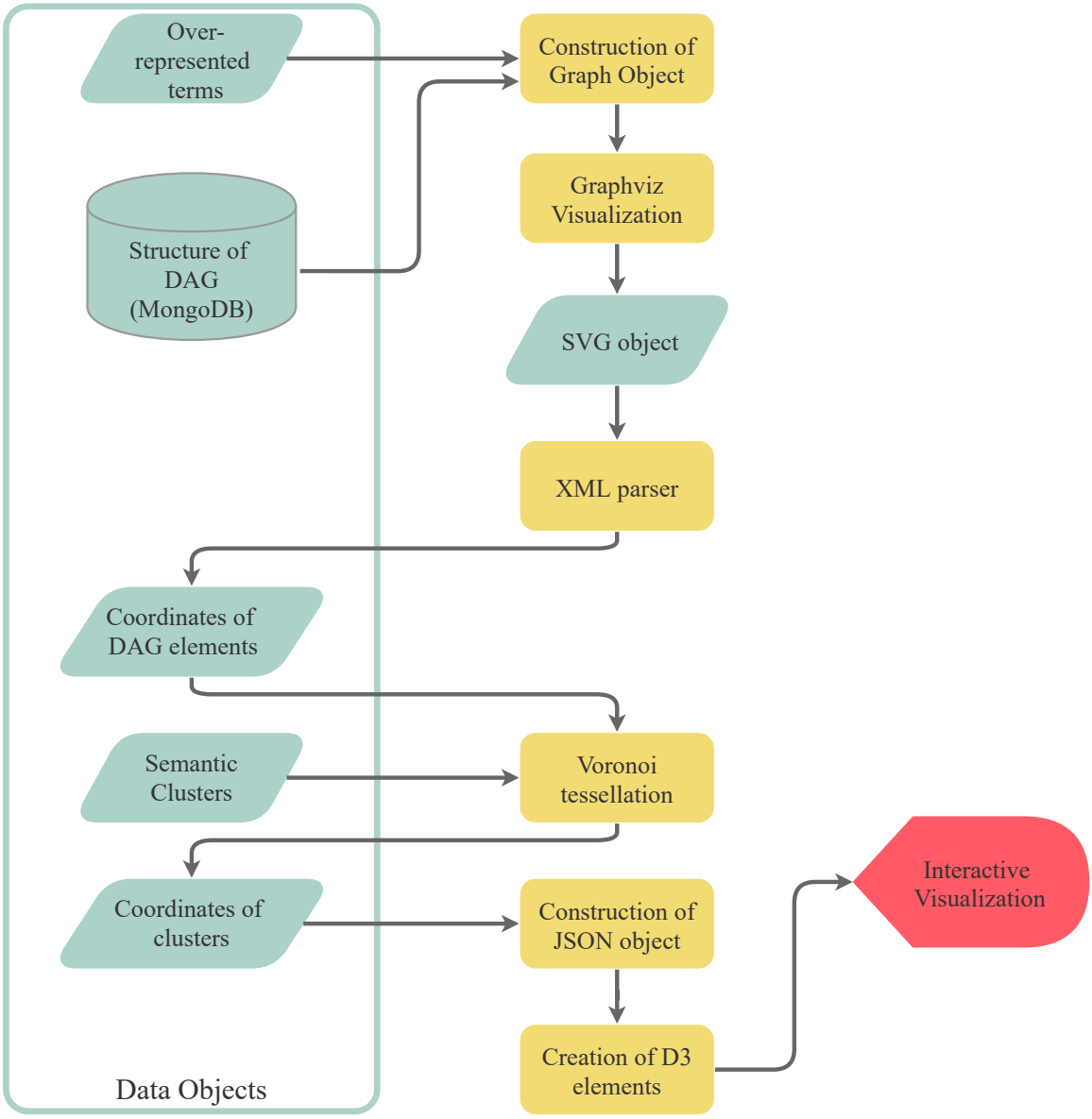


Figure 3-3: The workflow for data visualization. BioInfoMiner results are transformed in JSON data to be uploaded as SVG elements, using D3.js library, on a front-end framework.

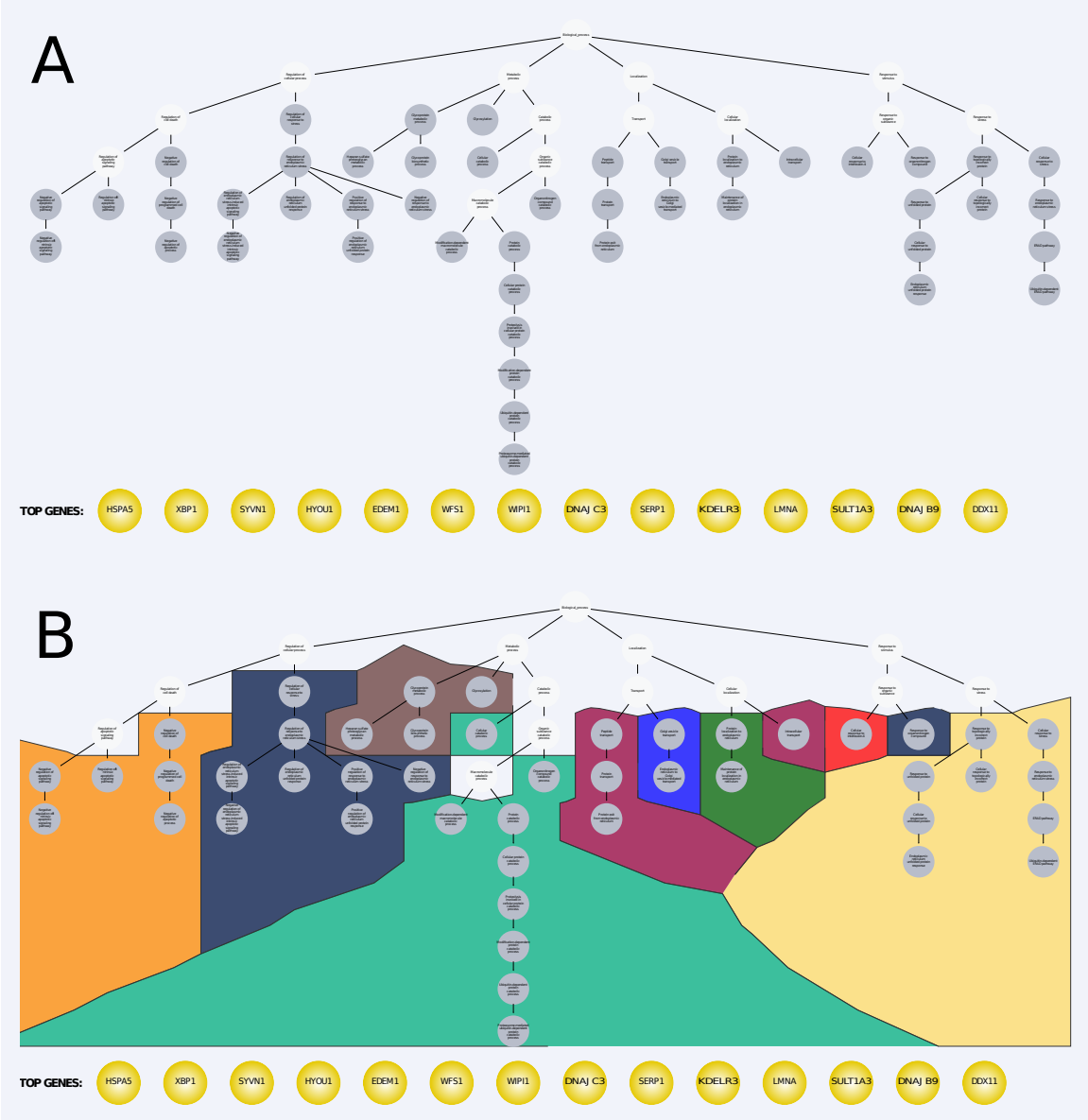


Figure 3-4: Visualization of BioInfoMiner results. (A) A pruned tree depicts the hierarchical organization of over-represented terms (grey nodes), while intermediate non-significant terms are condensed into single entities (white nodes). Below the area of the hierarchical tree, another plot displays the prioritized genes as circles. Interactive functions reveal their hidden associations, according to BioInfoMiner analysis. (B) Semantic clusters are visualized using the concept of Voronoi tessellation. Each prioritized gene is associated with various components on this topology.

3.3 Discussion

3.3.1 BioInfoMiner Novelties

The technical advancements of BioInfoMiner modules compared to the previous algorithms, developed in the laboratory of Metabolic Engineering and Bioinformatics Program of NHRF are presented in section 3.2.1. The current software architecture facilitates the integration of biological knowledge for different species, enabling the translational analysis of any organism which has been annotated in genome-wide scale. While the majority of pathway analysis and gene prioritization tools focus only on human and mouse genomes, the current version has been designed to analyze data for twenty eukaryotes (animals, fungi and plants; Table A.1). Also, the analysis is not limited to GO annotation, as various ontologies have been integrated (Plant Ontology, WormBase and Zebrafish Ontology) in the database. Hence, scientific teams which study solely a specific model organism (such as *Caenorhabditis elegans*) could use BioInfoMiner for the interpretation of their genome-wide experiments, exploiting species-specific knowledge.

In contrast to widely-used tools for pathway analysis, BioInfoMiner provides an automated solution to winnow not only the important pathways but also their underlying associations with nodal genes on their network. Also, the automated extraction of gene signature, without the usage of seed gene lists or keyword mappings, is an advantage comparing to other gene prioritization methods, described in section 1.5.2. The final signature constitutes a succinct yet informative and novel collection of genes, exclusively defined for the examined pathological state or phenotype. That procedure could be used as a feature selection method, which uses semantic and not experimental criteria to reduce the dimensionality of input vector. While the analysis does not provide any indication about the causal merit of prioritized genes, they could be considered as potential nodal genes to portray the studied phenotypic stratifications. The reason behind of this is that they bear a substantially differentiated profile in genome or trancriptome level in combination with a central position on the semantic topology. Additionally, the visualization workflow provides a powerful solution to integrate these data in a comprehensive map and untangle the genomic complexity, helping the user to understand how the emergent semantic networks are associated with the proposed gene signature. The interactive environment

provides extra navigation tools to display user-defined components of semantic network in conjunction with experimental and background information.

3.3.2 Future Development

Concerning the future work on BioInfoMiner methodologies, the process of pathway analysis could be refined in the level of inference. In general, one disadvantage of pathway analysis is that it requires the usage of arbitrary statistical thresholds in different steps. If the applied approach uses an over-representation statistical test, then it is necessary to predefine the list of differentially expressed genes (or gene products), using a p-value threshold (usually 0.05). Despite that some gene set and topology-based approaches developed to surpass that drawback, taking into account the whole p-values distribution, the subsequent meta-analysis adopts arbitrary thresholds to specify the set of significant pathways. In both cases, p-value serves as an indicator of biological signal which varies continuously between the background noise and the extreme signal prominences. However, neither the statistical model nor the strategies of multiple hypothesis testing could guarantee that extreme signals imply comparable biological association or that weak signals are not linked with moderate, yet interrelated with the examined condition, molecular events. Thus, the sharp dichotomization in statistically significant and not significant events, solely based on statistical criteria, could demote the quality of the extracted biological information and the plausibility of the subsequent interpretation.

A premature idea to overcome the issue of arbitrary statistical thresholds was presented at the 8th International Conference on Systems Biology (ISB) [159]. In that work, an entropy-based, free of statistical thresholds, workflow was constructed and applied in a DNA microarray dataset for muscular aging, targeting to associate the given phenotype with a broad, yet highly informative (with noise minimization) gene list. The proposed algorithm involved an iterative execution of StRAnGER on an incremental set of genes, according to their p-values ranking, and the partitioning of enriched GO terms into informative and potentially noisy, based on specific criteria. Following that process, the Shannon entropy [160] genes was estimated in each round, taking into consideration the associated GO terms. The main goal was to detect the maximal set of genes for which the measurement of the aggregated Shannon entropy reached a local minimum. The analysis revealed that genes with traditionally great p-values

(>0.05), in the level of null hypothesis acceptance, could incorporate important biological information, rather than noise, increasing the information content of reliable functional terms. That work could pave the way for the construction of an automated, relieved from statistical cutoffs, information-based approach, in order to improve the pathway analysis module. For that purpose, the theoretic framework of noise-chasing method and that of mathematical optimization need to be determined, while the final algorithm should be validated with plenty of benchmark datasets.

3.3.3 Conclusion

Overall, the computational workflow of BioInfoMiner serves as a data-driven, agnostic operator, able to deliver the systemic interpretation of the input list, regardless the omic approach and the experimental conditions from which it was derived. The gene list could have been derived either from modern high-throughput biological experiments or annotation procedures, where bio-curators relate manually or electronically an entity (disease, phenotypic trait, molecular mechanism, etc.) with a list of genes, under certain criteria. Chapters 6, 7 and 8 present how BioInfoMiner and its software library could be used in different scenarios to elucidate various questions through the implementation of semantic networks analysis.

Chapter 4

BioTranslator: A BioInfoMiner Instantiation for the Semantic Interpretation of Metagenomes

This chapter presents the BioTranslator tool, a Galaxy-integrated automated workflow, for the semantic interpretation of metagenomic samples. BioTranslator is an instance of BioInfoMiner, appropriately configured to analyze metagenomic datasets in the environment of Galaxy engine. It constitutes a part of ANASTASIA, a web-based platform and repository for the systemic analysis of metagenomic samples [161]. Initially it was developed in the framework of the European FP7 project HotZyme, whose aim was to perform exhaustive analysis of metagenomes derived from thermal springs around the globe and discover new enzymes of industrial interest. The current version of ANASTASIA is available at motherbox.chemeng.ntua.gr/anastasia_dev.

4.1 Introduction

Metagenomics is a research discipline in molecular biology, applied in environmental and clinical studies to explore the genomic and enzymatic content of microbial communities [162]. In environmental metagenomics [163], mixed biological communities from common environments (soil and seas) or ecosystems with utmost conditions (volcanoes, thermal springs and glaciers) are collected and analyzed without prior cultivation. The in-depth study of a community genotype (metagenome) aims to unravel not only the phylogenetic properties of

the environmental niche itself, but also the molecular mechanisms that provide adaptability and sustainability in such conditions. These mechanisms and the underlying network of interactions could comprise useful novel "recipes" for the biotechnological industry. Thus, metagenomics provides industry with an unprecedented chance to bring novel biomolecules into industrial application. Rather than designing processes that are able to maintain enzymes originally vulnerable to extreme conditions, it is preferable that naturally tolerant enzymes, regulated to persist in such conditions, could be used as an alternative and efficient tool for industrial procedures.

Metagenomic analysis is not limited to environmental samples. Understanding the contribution of microbiome regarding modulation of host response, is another highly-important, emergent, scientific field. The study of human microbiome aims to elucidate the role of microbial communities in human health and how modifications of microbial composition affect the healthy state, provoking specific diseases [164]. More than 10000 microbial species are hosted in and on the human body, outnumbering the total number of human cells by an estimated factor of 10 fold [165]. The composition of the microbiome varies by anatomical site, indicating the existence of specific evolutionary niches, based on co-adaptation [166]. The taxonomic profile of each anatomical site reflects the evolutionary adaptation of host and microbial cells, which shapes a complex cooperative network of functions and is characterized by long-term equilibrium and robustness to accidental perturbations [167]. However, that equilibrium is not stable, as it can be influenced by dietary changes, exposure to antibiotics and infections. These factors alter the microbiome composition, which could lead to new pathogenic states, affecting the overall human homeostasis. As a result, the in-depth exploration of human microbiome functionalities and how it becomes pathogenic is a crucial task for the comprehension and the treatment of related diseases.

Metagenome analysis workflows integrate a plethora of computational tasks. Samples acquisition and sequencing (i.e. the full documentation of the nucleotide sequences that constitute the metagenome) ends up to a chaotic collection of DNA sequence chunks (billions or trillions), which constitute the starting point (raw data) of computational analysis. Various procedures, such as i. quality control, ii. metagenome assembly, iii. gene detection, iv. gene annotation, v. taxonomic analysis and vi. comparative analysis, should be performed in order to elucidate those complex biological systems enigmas. For this scope, ANASTASIA,

a web-based platform and repository, utilizing the Galaxy workflow engine [168], was developed. ANASTASIA integrates a wide variety of bioinformatics tools for the systemic analysis of metagenomic samples, combining them to transparent workflows, executing composite computational tasks, at the push of a button. The motivation in this work was to provide an automated solution for the functional description of microbial communities. To that end, BioInfoMiner computational workflow (described in chapter 3) was configured and adapted to the demands of metagenomic data. That version was named BioTranslator and it was integrated in the ANASTASIA repository.

4.2 Methods

BioTranslator workflow is an instance of BioInfoMiner. BioInfoMiner was developed to analyze organism-specific omic data, whereas the main idea in this work was to build a non-specified methodology, sufficient to functionally interpret a metagenomic sample regardless its phylogenetic composition. GO was selected for that task, as it provides a universal functional annotation and thousands prokaryotic species, with biomedical or industrial interest, have been annotated in genome scale. To date, it is the foremost ontology for the massive annotation of prokaryotic world. Furthermore, as the final goal was to integrate BioTranslator in ANASTASIA platform and its automated pipelines, additional modules needed to be constructed. Taking the above under consideration, adjustments on GO annotation and computational workflow were carried out.

4.2.1 Adjustment of the GO Annotation

The processing of metagenomic raw data correlates parts of the detected open reading frames (ORFs) with known sequences or protein domains, based on sequence alignment algorithms. Sequence matches with high-similarity are used to annotate the genomic content of microbial community. An appropriate pathway analysis tool needs to exploit all the available functional annotation of the prokaryotic world to interpret these results. The UniProtKB/SwissProt database [169] includes manually curated descriptions for more than 350k known prokaryotic proteins. Thus, to overcome the existence of organism-specific databases, GO annotations of different species, retrieved from UniProtKB/SwissProt, were combined in a unified schema. All the orthologous genes (i.e. genes in different

species that origin from a common ancestral gene) were mapped together, combining their functional annotations to produce a unique gene – GO mapping. In order to eliminate the annotation bias of extensively studied prokaryotes, infrequent associations were filtered out. The relative frequency of each gene – GO term pair was calculated as the ratio of the species which are annotated with that pair to the whole set of organisms, which contain that gene in their genome. Gene-specific distributions of relative frequencies were calculated, so that each gene - GO term pair with value lower than the median of distribution was excluded from the final annotation set of that gene. The output versions of GO domains shape a global description of biological processes, cellular components and molecular functions that exist in the prokaryotic kingdom, regardless of the taxonomic details.

4.2.2 Adjustment of the Computational Workflow

The software library was extended to facilitate the integration of the computational workflow in automated analytical pipelines of ANASTASIA repository. The input data could be either a predefined list of genes or gene products, as in the original workflow, or a BLASTp [133] output (specific tabular format and executed on the SwissProt database), derived from previous analytical tasks (Fig. 4-1). In order to keep the most reliable BLAST hits, BioTranslator adopts strict alignment criteria, filtering out matches with query coverage lower than 90% or subject coverage lower than 50% and it accepts a user-defined threshold about the hits' e-value. The best hit of each query is kept and UniProt IDs are translated into the respective gene symbols. Regardless of the initial input, pathway analysis extracts a set of statistically significant terms, as they are described in the three GO domains. The user is able to determine the domain which will be used for the prioritization of genes. Hence, the second step uses the enriched part of that ontological graph to exploit its topology and disclose the most critical genes that could be assumed as the master regulators of the microbial community under study.

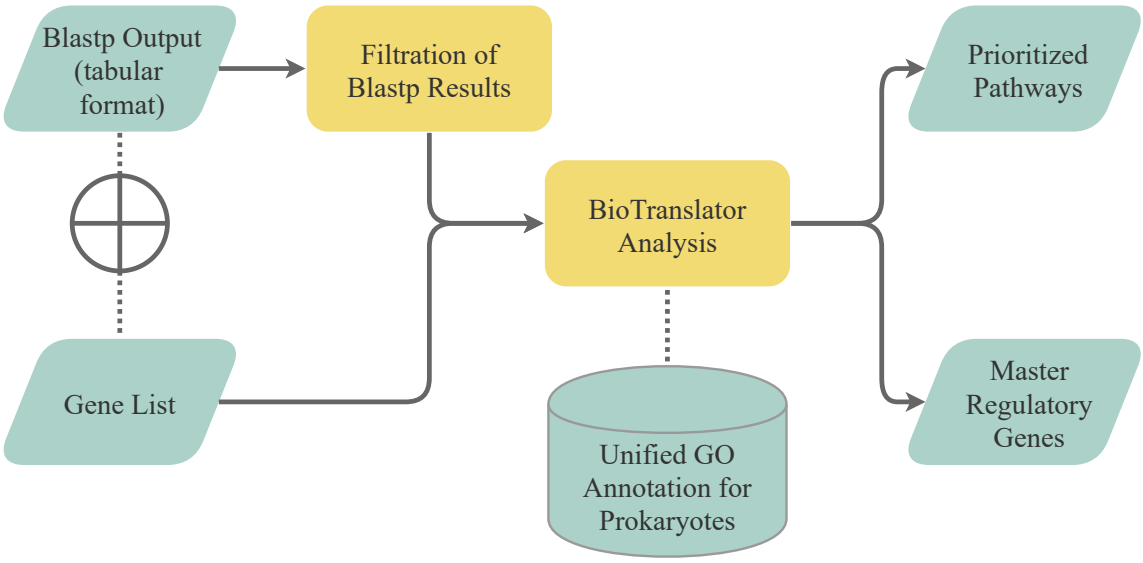


Figure 4-1: Computational workflow of BioTranslator. The user could start the analysis with a pre-defined gene list, or sequence similarities data derived from BLASTp analysis. BLASTp results are filtered to keep the most reliable matches. Pathway analysis and master regulators detection are based on BioTranslator core algorithms. The outputs are presented in the Galaxy front-end interface.

4.3 Discussion

4.3.1 Future Development

BioTranslator was developed for the automated interpretation of de novo assembled environmental metagenomic samples, however its future versions need to be adapted in the recent advances of metagenomic analysis and the demands of human microbiome studies. The unprecedented characterization of human microbiota has shifted the attention of scientific community to the identification of the causal relationships between host-associated microorganisms and various pathologies. Although the majority of microbial universe remains undefined, the intensive research have paved the way for the construction of microbial reference genomes and pan-genome models [170–172]. The exploitation of these repositories could facilitate the targeted functional analysis of specific species or genera, whereas the combined analysis of host omic data could reveal the crucial network of interactions and its underlying disparities in different pathological states. Gene prioritization step of BioTranslator could be used in such a scenario to glean the master regulators of these synergistic networks.

Metabolic reconstruction is another approach to explore the functional capacity of a (meta)genome, applied to design the connectivity structure of metabolism of the investigated organism or community [173, 174]. However metabolism is a part of cellular mechanisms and cannot be interpreted as the entire functional description. On the other hand, BioTranslator uses the GO annotation to delineate a global network of functional components and detect their regulatory markers. Ideally, both these methods could be used in parallel for the interpretation of a metagenomic sample. BioTranslator could be used to pinpoint the generic network of mechanisms, shaping the primary structure of the underlying functionality, while metabolic reconstruction could detect significant metabolic reactions and pathways, providing a composite description about the energy production process. MetaCyc[44] and KEGG [13] are the prominent databases for this purpose. The interconnection of such methodologies in a unified, integrated solution, could be a future project for the improvement of translational analysis of metagenomes.

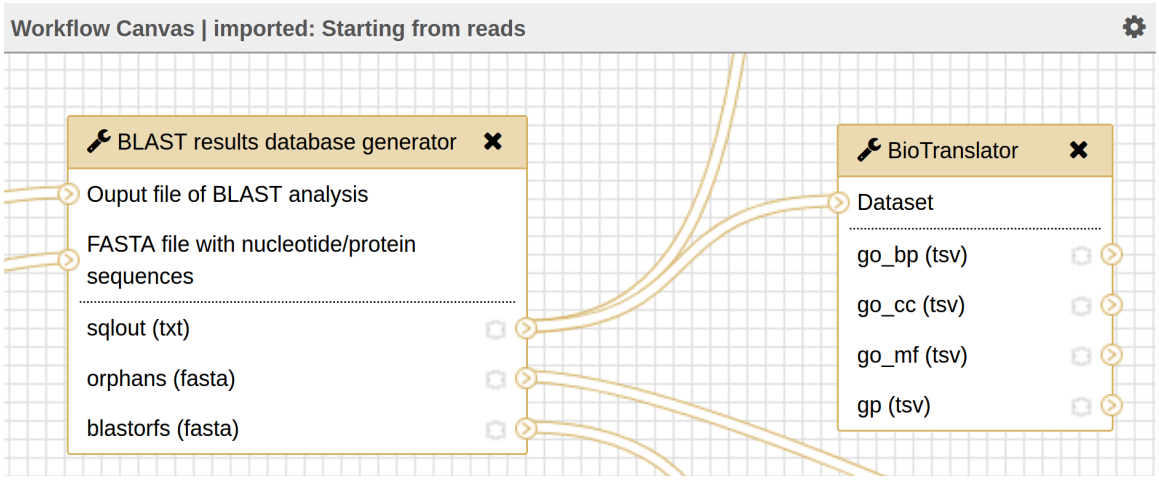


Figure 4-2: Part of ANASTASIA workflow for the analysis of metagenomic samples. BioTranslator module receives the output of BLASTp.

4.3.2 Conclusion

The integration of BioTranslator in the environment of Galaxy platform, demonstrated the portability of the main software library described in chapter 3. The computational workflow of BioInfoMiner can be easily embedded in integrative computational environments, as an operator. Also, its modular architecture

allowed the adjustment of background database in order to analyze a customized version of GO for prokaryotes. The configuration to accept BLASTp outputs facilitates its incorporation in the complex pipelines of ANASTASIA (Fig. 4-2). Thus, even the most inexperienced users could use it in the public server (motherbox.chemeng.ntua.gr/anastasia_dev) via the graphic user interface of Galaxy, as a standalone tool or built-in the automated pipelines.

Chapter 5

Qualitative Evaluation of Semantic Similarity Measures

Several measures have been developed to apply semantic analysis on ontological graphs. Some of them have been described in section 2.4, while they have been implemented as functions in the object-oriented software library of BioInfoMiner (section 3.2.4). The purpose of this study was to winnow the crucial factors that impinge on the semantic association of two terms and to evaluate existing similarity measures, in terms of their consistency with these factors. To address this, an artificial ontological tree was constructed and pairs of terms were ranked, based on human perception. Different measures were applied, to assess their ability to reproduce the same ranking. The results were presented in the Bio-Ontologies COSI (Communities of Special Interest) track of the ISMB/ECCB Conference in 2019 [175].

5.1 Introduction

The current information-theoretic measures take into consideration the topology of terms on the directed graph. The foremost concept that is used to determine the semantic specificity of a term is its Information Content (IC) (equation 2.1). Additional values have been proposed to characterize terms' topology (equation 2.3, 2.4), pointing out several limitations of IC. The respective measures assume that the adjacency of two terms is positively correlated with the topological position and the extent of their common ancestors set. Preliminary measures combined the idea of the most informative common ancestor (MICA; equation 2.5) and that

of IC to define the similarity of two terms. When all nodes but MICA are ignored, different possible interpretations of the semantic concepts are disregarded [141]. To overcome this limitation, different ancestor strategies have been constructed, considering either all common ancestors or only a part of them (equation 2.8, 2.10 and 2.11).

The unified theoretic framework of Mazandu and Mulder [143] postulates that all the measures borrow their formulas by Set Theory and the similarity between two terms could be expressed as the ratio of their semantic intersection and union. Intersection is determined by their Shared Information (SI), namely the topological specificity of the common ancestors. Likewise, union is represented as the sum of their semantic specificity. Analogizing the topological space of ontological terms as sets of objects, these assumptions become defective. Particularly, the effect of uncommon disjunctive ancestral paths is considered only when one of the aforementioned ancestor strategies is used to calculate the SI, instead of MICA. Also, the effect of terms' specificity is canceled when the intersection is empty, so the numerator is zero. Furthermore, that reasoning does not provide any explication about how the existence of common descendant terms could influence the similarity. Intuitively, two terms with common semantic sub-entities are probably more relevant than a disjoint pair. Thus, the existed measures address partly the distinct topological properties and pitfalls of ontologies, as they perform simplified semantic comparisons, ignoring a part of crucial features.

Taking into account all the aspects of topological complexity of biomedical ontologies, a more precise framework need to be developed. In this work, the theoretic concepts of semantic analysis were clarified and a set of important factors that need to be considered during the comparisons was proposed. The level of conformity of the existing measures to these factors was assessed, by constructing a ranking of term pairs on an artificial directed graph, based on human perception. The measures were examined for their ability to correctly prioritize these pairs, complying to the benchmark ranking. The following factors determine the theoretic framework, in order to define a more efficient semantic measure:

1. *Specificity of the common ancestors*: A pair of terms, whose common ancestors have high specificity, is more similar than a pair of terms, which have approximately the same degree of specificity with the first pair, but

markedly less informative common ancestors.

2. *Multiple parent inheritance*: The existence common disjunctive ancestors, instead of a unique common ancestry, does not decrease the similarity of two terms. In contrast, similarity is decreased due to the existence of uncommon disjunctive ancestors.

3. *Specificity of the compared terms*: General terms with low specificity are more similar than a pair of terms with high specificity, which have the same common ancestry paths - analogous with the concept of Lin and Aggregate IC (equation 2.13, 2.18).

4. *Common descendant terms*: The presence of common descendants indicates the existence of common semantic sub-entities (which potentially emerges from overlapping genomic annotations) and increases the semantic similarity of terms.

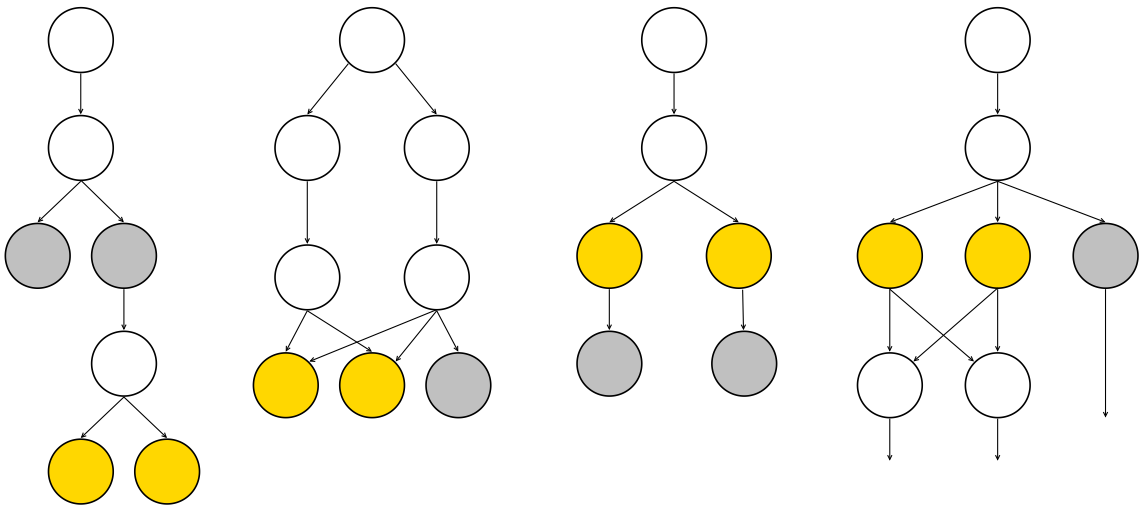


Figure 5-1: Illustration of the proposed factors on appropriate graphical instances. Concerning to factors 1 and 3, the gold coloured pair has greater semantic similarity than the grey one. Factors 2 and 4 state that the similarity between gold nodes surpasses their respective individual similarities with the grey node.

5.2 Methods

5.2.1 Benchmarking Using the Human Perception

In the past, numerous studies investigated the efficiency of semantic similarity measures [139, 142, 143, 148, 151, 176, 177]. Traditionally, the evaluation process uses a list of proteins and their functional annotation in order to calculate a similarity matrix for each measure. Then, an adjacency mapping of proteins is assumed as benchmark, using either a quantitative feature of proteins (sequence, structure or expression levels) or their proximities on protein-protein interactions (PPI) network. In the first case, a similarity matrix is constructed and the ability of each measure to reproduce relevant similarities is examined. Using the PPI network, an accurate measure is expected to infer high similarity values to interacting proteins, or proteins which act in the same pathways. However, these features could not entirely guarantee functional coherence and therefore no gold standard datasets have been developed (each author devises a different case study).

Alternatively, some studies [139, 142] have referred that semantic comparison should be consistent with the inferences of human perception. Namely, the deductive logic could be used to evaluate isolated, non-complex scenarios, investigating the accuracy of measures based only on one independent variable. That method was used in the present study to estimate the efficiency of measures described in section 2.4. That strategy needs the usage of a benchmark ontological graph, free of the common pitfalls of biomedical ontologies (e.g. edges with various semantic weights, deep and shallow branches, high degree of interconnections). Hence, an artificial ontology was constructed to evaluate the semantic similarity measures (Fig. 5-2 A). It was manually designed to have four distinct levels of specificity (IC values) and not complex interconnections, in order to easily find pairs of terms with evident difference in semantic relatedness, based on the first factor. Twelve pairs of terms with considerably different similarities based on the proposed system of rules were prioritized (Fig. 5-2 B). The following paragraph describes the thought process for that prioritization.

Pairs (D1,D2) and (D3,D4) have the highest similarity as their MICA terms belong to the C-level (graph tier with the most informative parents) and none of these nodes has any disjoint ancestral path. That issue is observed in (D3,D5) were D5 has two separate parents. As a result, it has lower similarity compared

to the previous two pairs. Concerning to pairs which have their MICA in the B-level, (C1,C2) pair has greater similarity than (C3,C4) due to the set of common descendants. This indicates an internal semantic overlap of C1 and C2, setting them closer on the topological space. (A1,C2) is the only pair with MICA in the A-level, while their similarity is increased due to their indirect relationship. All the other pairs have the root as MICA, so their SI value is the minimum. However, they are prioritized according to the last two proposed factors, which suggest that the existence of common descendants and the low semantic specificity are positively correlated with the increase of semantic similarity.

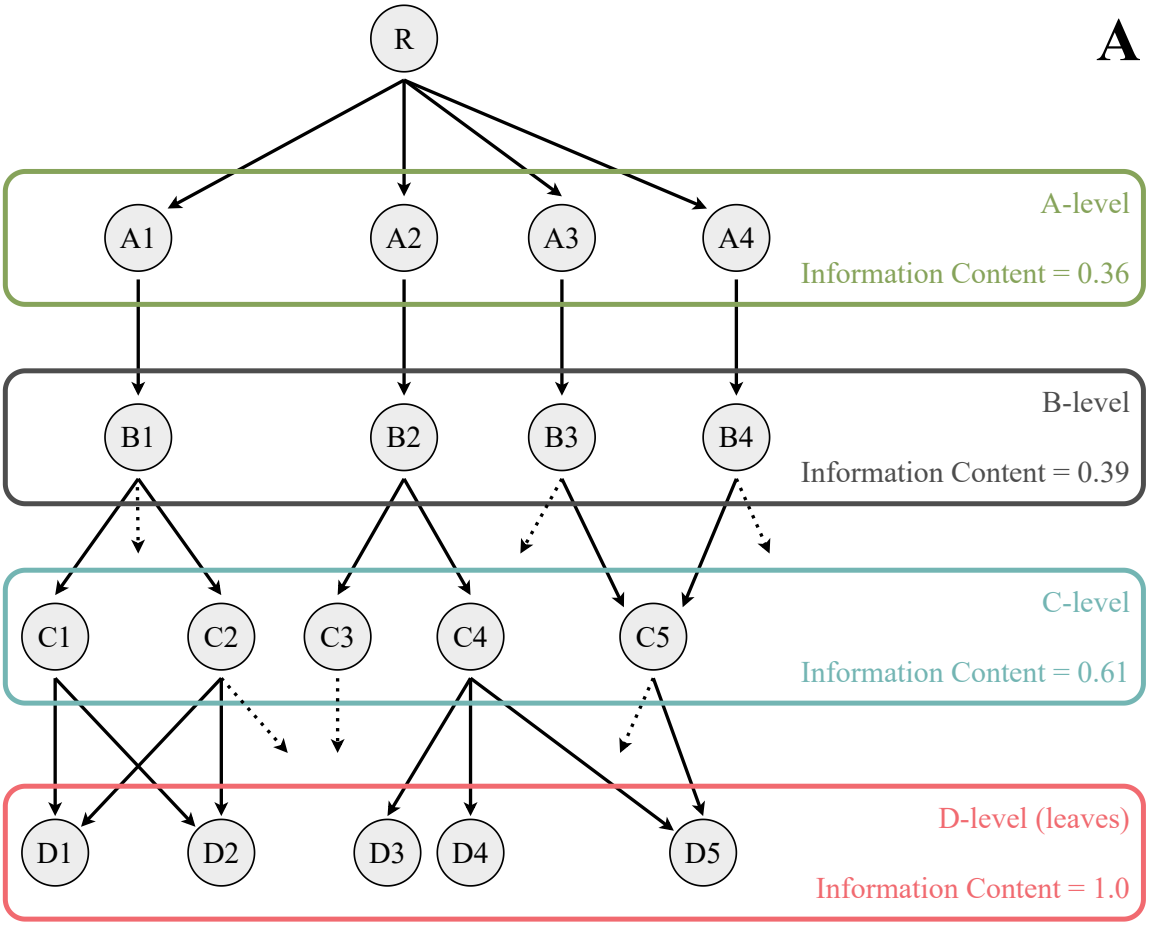
5.2.2 Evaluation Metrics

All the possible combinations of semantic similarity measures and SI strategies were implemented to prioritize the above pairs. The congruity of each sorted list with the benchmark was evaluated based on two criteria: the Manhattan distance [178] and the Kendall's tau-b correlation coefficient [179], which measures the monotone relationship between two ranks. Given a set of N elements $\{e_1, e_2, ..., e_n\}$ and two ranking vectors of them, $X = [x_1, ..., x_n]$ and $Y = [y_1, ..., y_n]$, where x_i, y_j are the rankings of element e_i in X and Y, then the Manhattan distance of X and Y is defined as:

$$MD(X,Y) = \sum_{i \in [1,n]}^{N} |x_i - y_i| \tag{5.1}$$

The Kendall's tau-b correlation coefficient examines all the $\binom{N}{2}$ combinations of pairs (x_i, y_i) and (x_j, y_j), where $i < j$: $\big[(x_1, y_1), (x_2, y_2)\big], ..., \big[(x_{n-1}, y_{n-1}), (x_n, y_n)\big]$. The number of concordant pairs (namely either both $x_1 > x_2$ and $y_1 > y_2$ hold or both $x_1 < x_2$ and $y_1 < y_2$), and the number of discordant pairs are counted in order to calculate the coefficient:

$$Kendall(X,Y) = \frac{\big|\text{concordant pairs}\big| - \big|\text{discordant pairs}\big|}{\frac{N \times (N-1)}{2}} \tag{5.2}$$



B

Rank	Pair	Ranking Explanation
1	(D3, D4)	MICA in C-level
1	(D1, D2)	Multiple common parents in C-level & no uncommon ancestors
2	(D3, D5)	MICA in C-level and one uncommon ancestry path
3	(C1, C2)	MICA in B-level & common descendants
4	(C3, C4)	MICA in B-level
5	(A1, C2)	MICA in A-level, ancestor-descendant relation
6	(B3, B4)	R as MICA, B-level terms, common child
7	(A3, A2)	R as MICA, A-level terms, remote common descendant
8	(A1, A2)	R as MICA, A-level terms
9	(B1, B2)	R as MICA, B-level terms
10	(A1, D5)	R as MICA, no correlation
11	(D1, D5)	R as MICA, no correlation

Figure 5-2: An artificial ontology which was constructed to evaluate the semantic similarity measures. (A) Its structure consists of four distinct levels of specificity - terms of the same level have the same IC value. (B) Twelve pairs were prioritized, as their semantic similarity was qualitatively estimated using the proposed factors.

5.3 Results & Discussion

5.3.1 Comparison of Semantic Measures

None of the existed approaches could infallibly reproduce the reference prioritization (Table 5.1). Aggregate IC, which considers the whole ancestral sub-graph of each term specificity calculation, had the best performance. Schlicker and Resnik measures with XGraSM and DiShIn strategies showed also adequate accuracy. They both adopt a conservative logic, producing a lot of similarly equal pairs, especially in the upper levels of the graph. On the other hand, Jiang and Conrath measure had the worst performance. Overall, the adoption of DiShIn method to calculate the similarity of two terms based on their common disjunctive ancestors, instead of MICA, increases the performance of measures, while the XGraSM technique improves only the performance of Schlicker and Resnik measures. Also, it is evident that none of the measures considers the existence of common downstream sub-graphs between two terms.

The main difference between Aggregate IC and the other measures is that it succeeded to assign different similarity scores for the majority of pairs, especially in the tail of the sorted list. The reason for that behavior is that Aggregate IC formula uses the property of SV, instead of IC. By definition SV is greater than zero (equation 2.3), so the semantic similarity of a pair with the graph root as MICA is nonzero and it is determined by terms' specificity. In contrast, all the other measures use the IC, which is zero for the root term and therefore, all the uncorrelated terms have similarity equal to zero. Despite that Aggregate IC quantifies the contribution of the first three proposed factors and surpasses all the other measures, it produces discrepancies with the reference ranking for adjacent pairs.

Inconsistencies between the Manhattan distance and Kendall's correlation are observed comparing the strategy of MICA with those of XGraSM and DiShIn. In general, the complex methods interpret in a superior way the first three proposed factors (lower Manhattan distance values) comparing to MICA. However, they have slightly lower Kendall's correlation values. The reason for that contradiction is that they are more sensitive to produce different tiers of similarity scores on the graph which could lead to disconcordances with the reference ranking. On the other hand, MICA generates distinct tiers, where a lot of pairs have the same similarity and they are prioritized equally. These ranking ties are not considered

Table 5.1: Results of the Evaluation of Semantic Similarity Measures

Measure	Strategy of SI	Manhattan Distance	Kendall Tau-b Correlation	Ranking
Aggregate IC	-	18	0.719	[1,2,4,3,3,7,6,5,5,6,8,9]
Resnik	DiShin	28	0.762	[1,1,4,2,2,3,5,5,5,5,5,5]
Schlicker	DiShin	28	0.762	[1,1,4,2,2,3,5,5,5,5,5,5]
Resnik	XGraSM	30	0.794	[1,2,2,3,3,4,4,4,4,4,4,4]
Schlicker	XGraSM	30	0.794	[1,2,2,3,3,4,4,4,4,4,4,4]
Lin	DiShin	32	0.620	[2,2,4,1,1,3,5,5,5,5,5,5]
Nunivers	DiShin	32	0.620	[2,2,4,1,1,3,5,5,5,5,5,5]
Faith	DiShin	32	0.620	[2,2,4,1,1,3,5,5,5,5,5,5]
Resnik	MICA	33	0.850	[1,1,1,2,2,3,4,4,4,4,4,4]
Schlicker	MICA	33	0.850	[1,1,1,2,2,3,4,4,4,4,4,4]
Nunivers	MICA	36	0.633	[2,2,2,1,1,3,4,4,4,4,4,4]
Faith	MICA	36	0.633	[2,2,2,1,1,3,4,4,4,4,4,4]
Lin	MICA	36	0.633	[2,2,2,1,1,3,4,4,4,4,4,4]
Lin	XGraSM	37	0.567	[2,3,3,1,1,4,4,4,4,4,4,4]
Faith	XGraSM	37	0.567	[2,3,3,1,1,4,4,4,4,4,4,4]
Nunivers	XGraSM	37	0.567	[2,3,3,1,1,4,4,4,4,4,4,4]
Jiang & Conrath	MICA	42	0.338	[4,4,4,1,1,2,4,3,3,4,5,6]
Jiang & Conrath	DiShin	43	0.293	[4,4,6,1,1,2,4,3,3,4,5,7]
Jiang & Conrath	XGraSM	44	-0.064	[4,5,5,3,3,6,2,1,1,2,6,7]

as disconcordances from Kendall's equation and MICA produces better results for the same measure. In the light of the above, there are two different categories of measures, based on the causality of their inconsistencies. MICA tends to increase the bias of the semantic metric, whereas the other methods increase its variance. Hence, the selection between the complex ancestors strategies and MICA could be translated as a bias-variance trade-off.

5.3.2 Conclusion

To conclude, the evaluation revealed the advantages of some approaches, but also a general inadequacy of the existing methods to correctly assess all the

variables that influence the similarity of terms. To further explain the above results, Set Theory could be used to describe the structure of an ontology, while its topological space could be represented as an area-proportional Euler diagram. A similar idea has been presented in [180] and a toy model is depicted in Fig. 5-3. The root of DAG is illustrated as the superset which includes all the other terms. Its gradual segregation likens the semantic hereditary which traverses the graph, from root (set A) to leaves (sets E, F, G, H, I and J). The size of a set is inversely correlated to its semantic specificity and the similarity of two sets could be assessed by their contour distance. In such a way, sets I and J (descendants of C) are more similar than E and I (descendants of A), as they are conjointly located in a smaller area (1st factor). The existence of multiple parent inheritance moves a set away (e.g. H) from others (E, F and G), which are subsets of the same superset (2nd factor). The distance of two broad semantic entities (e.g. B and C) is smaller than that of their subsets (e.g. H and D), as their contours are de facto closer (3rd factor). Two sets with non-empty intersection (E and F) are located closer than others (E and G) in the same superset (4th factor). Such a transformation verifies the importance of the proposed factors and unveils the individual deficiencies of the existed measures, as they fulfill only a part of these factors. To improve the reliability of semantic analysis, an advanced measure or methodology needs to quantify the contribution of the proposed factors and integrate them as coefficients.

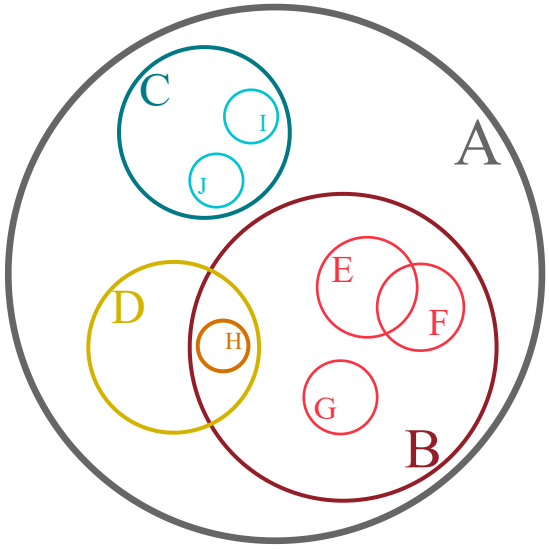


Figure 5-3: Visualization of ontological structure using the Euler diagram. The ontological graph could be presented as a collection of sets. Each set refers to a specific term and it is located in its superset or the intersection of its disjoint parental terms. This presentation elucidates the importance of the proposed factors in the calculation of semantic similarity.

Chapter 6

Case Study 1: Profiling of TRAIL-Induced Modulation of NK cells During Viral Infection

This chapter describes the transcriptomic and functional analysis of splenic NK cells profile, within the frames of a research study about the non-apoptotic role of TRAIL (TNF-related apoptosis-inducing ligand) in the regulation of NK cells activity during viral infection. BioInfoMiner workflow and software library were used to interpret the results of RNA-seq data processing and further elucidate the differences of NK cells response in presence and absence of TRAIL. The project was elaborated in terms of a six-month secondment in the Institute of Pathology of the University of Bern. The results have been published in [181].

6.1 Introduction

Natural killer (NK) cells are lymphoid cells which belong to the innate immune system and play a pivotal role in controlling viral infections and tumors. Their activation depends upon the relative contribution of extracellular signals to a variety of receptors on their surface. These receptors are divided in two main categories, concerning their effect: inhibitory and activating [182]. Inhibitory receptors suppress NK cells activation. Some of them interact with molecules of histocompatibility complex (MHC) class I, which are expressed in all nucleated cell types. Activating receptors work cooperatively and their combined signal has to surpass that of inhibitory receptors in order to trigger NK cells activation.

Virus-infected and tumour cells often present downregulated levels of MHC class I molecules, while cellular stress processes, such as DNA damage response, senescence program or tumor suppression cause the upregulation of ligands for activating receptors [183]. The combination of these molecular events incite NK cells to produce cytotoxic molecules in order to kill the transformed cells.

NK cells could trigger apoptosis secreting either cytotoxic granules or inflammatory cytokines [184]. The produced cytotoxic granules contain perforins, which form pores in the membrane of target cells, and granzymes (A or B), a type of serine proteases, which are released through perforin pores in the cytoplasm of target cells to induce lysis [185]. Also, NK cells produce various cytokines to stimulate the death of infected and malignant cells. Death receptor ligands, such as FasL, tumor necrosis factor (TNF-α/β) and TNF-related apoptosis-inducing ligand (designated as TRAIL and TNFSF10) bind on the cognate receptors of target cells and activate the extrinsic apoptotic pathway. IFN-γ is another cytokine whose production contributes to the development of innate and adaptive responses, through various molecular processes [186].

TRAIL functionality is ambiguous, as it does not solely trigger cell apoptosis. It is expressed on the surface of various immune system cells (T, NK cells, macrophages) in order to be anchored upon target cells. TRAIL could ligate on four different receptors in human cells, whereas two of them are death receptors (TRAIL-R1, TRAIL-R2) and the other two are antagonistic, that is they do not transmit signals to induce cell death (TRAIL-R3, TRAIL-R4). In mouse there is only one death receptor (DR5) and two potential decoy receptors (DCR1, DCR2) [187, 188]. Binding on death receptors, the death-inducing signaling complex (DISC) is formed, which subsequently activates downstream apoptotic cascades, leading to cellular death. On the other hand, decoy receptors as well other early molecular events [189] could lead to a non-canocical TRAIL signaling, which possibly inhibit cell apoptosis. The role of TRAIL in cancer has been studied extensively, as it constitutes a prominent target in cancer treatment due to its ability to mediate apoptosis without harming healthy tissues [190]. However, numerous studies have shown that the non-apoptotic TRAIL-induced signaling pathway does not only impedes cell death, but also elicits several pro-tumorigenic effects, contributing to cancer proliferation, invasion, migration and metastasis [189, 191–196]. Similarly, the role of TRAIL in viral infections is controversial, since many studies revealed impaired immune response due to the functionality of TRAIL. However, the underlying non-apoptotic mechanisms

remain uncertain. Cardoso et al [181] investigated the contribution of TRAIL to the immune response induced by lymphocytic choriomeningitis virus (LCMV) and the ability of TRAIL to activate non-apoptotic signaling pathways which hamper virus control.

The experimental design included four different mice populations, combining wild-type (WT) and TRAIL-deficient (TRAIL-/-, Knock Out-KO) genotypes with naive and LCMV infected states. Initial experiments pinpointed non-apoptotic functionalities of TRAIL during the infection. Specifically, TRAIL was associated with the regulation of CD8-positive T-cells response, impeding their activation and subsequently hindering virus clearance. Additional findings indicated that TRAIL restricts the production of cytokines in NK cells, as TRAIL-/- NK cells produced higher levels of IFN-γ. In contrast, NK cells of WT mice presented increased production of granzyme B molecules during infection. Granzyme B production in NK cells is induced by the IL-15 signaling cascade, which includes the PI3K-AKT-mTOR pathway. More precise experiments disclosed that TRAIL molecule positively regulates the cascade of IL-15 signaling to finally produce granzyme B. Afterwards, transcriptomic analysis was performed on isolated NK cells from naive and LCMV infected (24 hours) WT and TRAIL-deficient mice in order to validate the consistency of the aforementioned results and detect mechanistic differences between the respective immune system responses.

6.2 Methods

6.2.1 Transcriptome Profiling

Single-cell suspensions were prepared from spleens isolated from naive or from LCMV-WE-infected WT or TRAIL-deficient mice, 1 day post infection. NK cells were sort-purified by flow cytometry and re-suspended in TRI-reagent (Sigma-Aldrich). Barcoded stranded mRNA sequencing libraries were prepared from high-quality total RNA samples. Obtained libraries that passed the quality check step were pooled in equimolar amounts, and 1.8 pM solution of this pool was loaded on the Illumina sequencer NextSeq 500. Library preparation and sequencing was performed at the EMBL Genomics Core Facilities (GeneCore, Heidelberg, Germany). The final dataset included 14 samples of single-end, reverse-stranded reads, with average length 85bp. Their classification comprised of the pairwise combinations of two molecular conditions and two pathological

states (four classes) (Table 6.1).

Table 6.1: Classification of Samples

	Wild Type (WT)	Knock Out (KO)
Non Infected	01, 02, 03	04, 05, 06
Infected	07, 08, 09, 10	11, 12, 13, 14

6.2.2 Analytical Pipeline

RNA sequencing (RNA-seq) data processing was performed on the SevenBridges platform [197]. FastQC [198] was used to describe the average quality of reads for each sample. The first step of raw data processing was the alignment of reads to the reference genome. STAR aligner [199] was used for that task, as it achieves highly efficient mapping, combined with high execution speed. Mouse reference genome was retrieved from the Ensembl database [128]. In order to derive the appropriate data format for the differential expression analysis, HTSeq-count [200] was used to count the aligned reads of each exon.

Differential expression analysis was performed with DESeq2 [201]. Initially, four statistical comparisons were executed to investigate the differences of transcriptomic profiles: 1. WT non-infected vs WT infected, 2. KO non-infected vs KO infected, 3. Non-Infected WT vs Non-Infected KO and 4. Infected WT vs Infected KO. Considering that transcripts with adjusted p-value < 0.01 and absolute log2 fold change ≥ 2 were determined as differentiated, only the comparisons of WT and KO infected populations with the respective non-infected revealed significant disparities on expression profiles. Hence the analysis was limited to these two cases, which are referred as WT-activation (comparison no 1) and KO-activation (comparison no 2) hereinafter. The extracted lists of differentially expressed genes were used for the pathway analysis.

BioInfoMiner workflow (chapter 3) was applied for the functional interpretation of differentially expressed gene lists, using the Biological Process domain of GO and the Reactome database. The analysis led to a list of significant biological processes (GO terms) and a list of molecular mechanisms and pathways (Reactome terms) for each activation (WT and KO). Targeting to detect differences between these term sets, a semantic operator was developed to compare the lists of each ontology, based on the underlying graphical topology. Juxtaposing the enriched sets of each ontology on the ontological graph and exploiting

the ancestor-descendant relationships among them, the analysis revealed the biological processes and molecular pathways that are differentiated only in WT-activation or KO-activation. The difference between a set of semantic terms N to a set M was defined as:

$$Difference(N, M) = \left\{ t_i | t_i \in N \wedge t_i \notin M \wedge \left(d_j \notin M \forall d_j \in descendants(t_i) \right) \right\} \quad (6.1)$$

meaning that $Difference(N, M)$ is a subset of N, whose elements as well their descendants are not included in M. Thereby, pathway analysis and the subsequent semantic-based comparison of WT and KO activation disclosed precise uniquely-associated biological processes and mechanisms for each immune system response.

6.3 Results & Discussion

6.3.1 Mechanistic Modulations Correlated with TRAIL

As it mentioned in Methods section, differential expression analysis detected negligible differences between the WT and KO classes of each pathological state, thus these cases were excluded from the subsequent translational analysis. On the other hand, the comparisons of WT and KO infected populations with the respective non-infected samples (WT-activation and KO-activation) disclosed important differentiation in gene expression levels, notwithstanding that the isolation of cells was performed 24h after the infection. WT-activation and KO-activation were annotated by 1168 and 1356 differentially expressed genes respectively, while their intersection was 867 genes. Such an overlapping proportion indicates the consistency of immune response and the fact that TRAIL abrogation influences solely a part of immune system components.

The results of pathway analysis for GO and Reactome ontologies are presented in Table B.1, B.2 for WT-activation and in Table B.3, B.4 for KO-activation. The highly ranked terms assert the activation of immune response in both molecular conditions under the viral infection. Indicatively, both activations are annotated with terms related to response to virus, defense response, (innate, adaptive) immune system response, regulation of cytokine production, cytokine signaling pathways and regulation of cell death. The majority of these terms are enriched

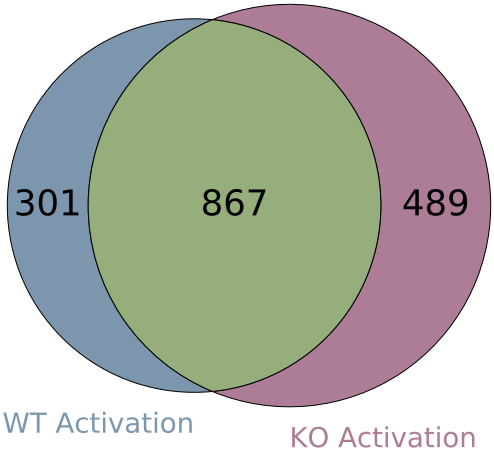


Figure 6-1: The Venn diagram shows the difference and overlap of genes that are differentially expressed in NK cells of WT versus Trail–/– mice during LCMV infection (for an adjusted p-value < 0.01 and absolute log2 fold change ≥ 2).

on a great scale, implying that an important proportion of differentiated genes has synergistic relations in the main component of NK cells activation.

The uniquely-associated biological processes and mechanisms for each activation have been colored in the aforementioned tables. Also, their graphical hierarchies are presented in the following figures (Fig. 6-2, 6-3, 6-4, 6-4). Of note, these sets corroborate a part of experimental evidences about the non-apoptotic functionality of TRAIL in NK cells activation. The suggestion that TRAIL modulates negatively the production of cytokines in NK cells during the infection, as the levels of IFN-γ found increased in Trail-/- NK cells, is verified in both WT and KO-activation GO term sets. WT-activation set contains the ”regulation of T cell cytokine production” and ”negative regulation of cytokine production involved in immune response” terms, whereas that of KO-activation includes the ”positive regulation of cytokine-mediated signaling pathway” and ”positive regulation of interferone-alpha production”, indicating the inverse correlation of TRAIL activation and cytokine production in NK cells. Additionally, the involvement of TRAIL for the downstream of IL-15/IL-15R signaling and granzyme B production is substantiated by Reactome pathways related to PI3K/AKT signaling and IL-2 family signaling (to which IL-15 belongs), which are differently affected in WT-activation. Conversely, KO-activation is annotated with the ”negative regulation of the PI3K/AKT network” and ”PI5P, PP2A, and IER3 regulate PI3K/AKT signaling”, in accordance with the experimental finding of diminished AKT phosphorylation and reduced granzyme B levels in TRAIL-deficient NK

cells.

Gene prioritization divulged 42 and 33 central-role genes for WT and KO activations respectively (Table 6.2). The intersection of them is composed by 22 genes, especially cytokines (interleukins, interferons, chemokines and a tumor necrosis factor), which are crucial molecules for various components of immune system response. IFN-γ is prioritized in both activations with significant degree of up-regulation, despite the result of TRAIL-modulated restriction of cytokines production. That was caused probably due to the existence of the related mRNAs into cell at the moment of NK cells isolation, despite the fact that previous proteomic experiments shown the blockage of IFN-γ in WT-activation. Interestingly, the gene set related to KO-activation includes a specific group of chemokines and more genes associated with chemotaxis comparing to that of WT-activation. That evidence is concordant with the existence of a distinct branch on the enriched GO ontological tree of KO-activation, which is associated with cell migration and chemotaxis. This study did not focus on these mechanisms, however future experiments could be performed to elucidate the influence of TRAIL deficiency in chemokines upregulation during viral infection and the role of subsequent chemotactic events in modulating virus clearance.

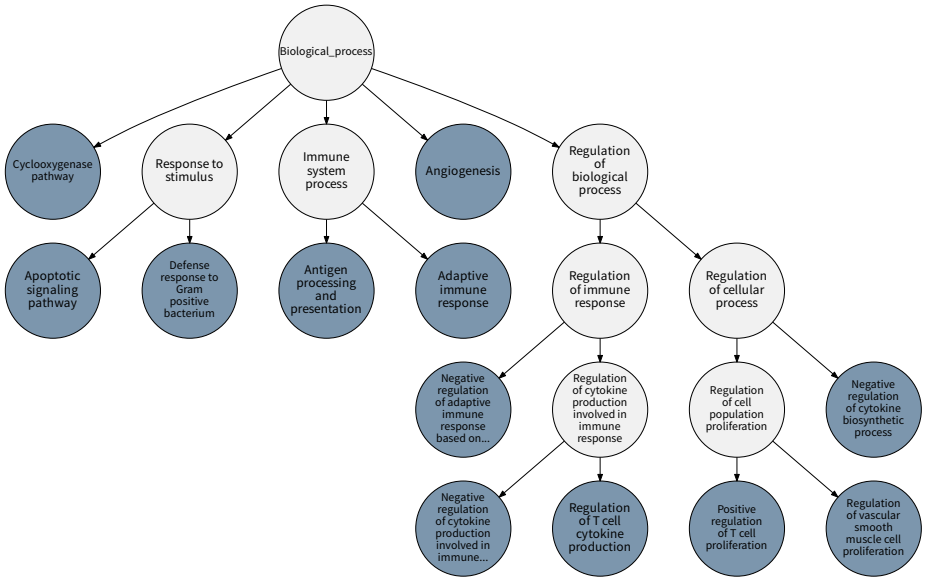


Figure 6-2: Part of the enriched GO terms for the WT-activation, clustered hierarchically in a succinct version of GO graph. The colored nodes depict terms uniquely enriched under WT-activation, defined using the equation equation 6.1.

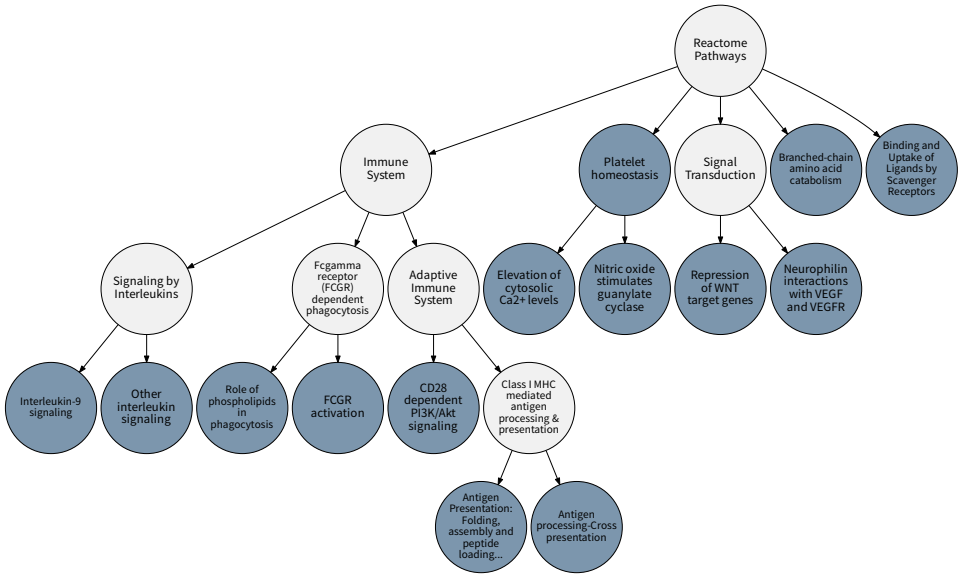


Figure 6-3: Part of the enriched Reactome terms for the WT-activation, clustered hierarchically in a succinct version of Reactome graph. The colored nodes depict terms uniquely enriched under WT-activation, defined using the equation equation 6.1.

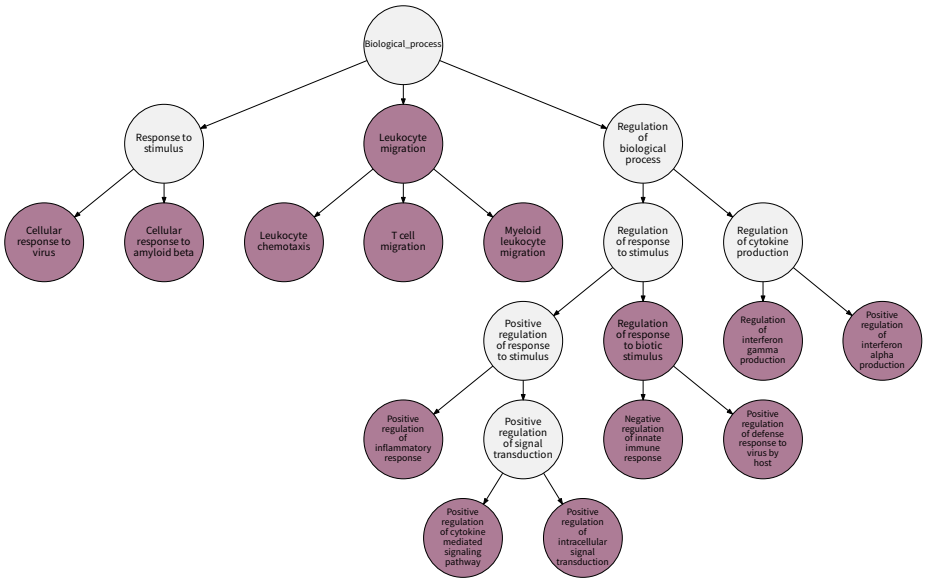


Figure 6-4: Part of the enriched GO terms for the KO-activation, clustered hierarchically in a succinct version of GO graph. The colored nodes depict terms uniquely enriched under KO-activation, defined using the equation equation 6.1.

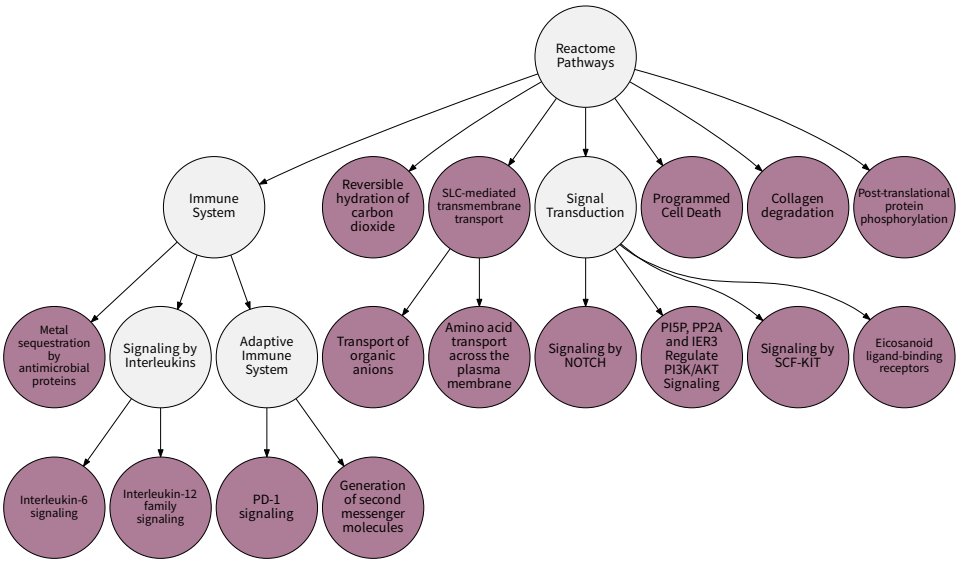


Figure 6-5: Part of the enriched Reactome terms for the KO-activation, clustered hierarchically in a succinct version of Reactome graph. The colored nodes depict terms uniquely enriched under KO-activation, defined using the equation equation 6.1.

Table 6.2: Prioritized Genes based on GO-BP under WT & KO Activations

The Union of Prioritized Genes			
Gene Symbol	Definition	Log2(FC) in WT-activation	Log2(FC) in KO-activation
Adar	Adenosine deaminase & RNA-specific	2.19	2.08
Bst2	Bone marrow stromal cell antigen 2	5.93	5.29
Ccl12	Chemokine (C-C motif) ligand 12	7.1	6.94
Ccl2	Chemokine (C-C motif) ligand 2	2.71	2.24
Cd36	CD36 molecule	-2.22	-2.63
Eif2ak2	Eukaryotic translation initiation factor 2-alpha kinase 2	3.74	3.6
Havcr2	Hepatitis A virus cellular receptor 2	3.07	2.44
Hfe	Hemochromatosis	-5.19	-4.15
Ifitm3	Interferon induced transmembrane protein 3	3.79	3.06
Ifnb1	Interferon beta 1 & fibroblast	7.02	6.29
Ifng	Interferon gamma	3.52	3.33
Il10	Interleukin 10	4.33	5.19
Il6	Interleukin 6	6.15	3.99
Irgm2	Immunity-related GTPase family M member 2	2.73	2.12
Lef1	Lymphoid enhancer binding factor 1	-3.13	-3.16
Plscr1	Phospholipid scramblase 1	2.58	2.61

Pml	Promyelocytic leukemia	3.3	3.28
Rsad2	Radical S-adenosyl methionine domain containing 2	6.75	6.98
Stat1	Signal transducer and activator of transcription 1	2.64	2.55
Tlr3	Toll-like receptor 3	3.99	3.02
Tnfsf4	Tumor necrosis factor (ligand) superfamily & member 4	4.13	4.14
Xcl1	Chemokine (C motif) ligand 1	2.46	2.8

Prioritized Genes Only for the WT-Activation

Gene Symbol	Definition	Log2(FC)
Bcl3	B cell leukemia/lymphoma 3	2.87
Casp3	Caspase 3	2.21
Cd3e	CD3 antigen & epsilon polypeptide	-5.34
Cd6	CD6 antigen	-3.63
Cd86	CD86 antigen	2.22
Fas	Fas (TNF receptor superfamily member 6)	2.91
Foxj1	Forkhead box J1	4.17
Foxp3	Forkhead box P3	-3.58
H2-T23	Histocompatibility 2 & T region locus 23	2.36
Ifna1	Interferon alpha 1	6.8
Igf1	Insulin-like growth factor 1	-4.64
Il12rb1	Interleukin 12 receptor & beta 1	3.02
Il2ra	Interleukin 2 receptor & alpha chain	5.6
Myc	Myelocytomatosis oncogene	2.23
P2rx7	Purinergic receptor P2X & ligand-gated ion channel, 7	-2.48
Ptgs2	Prostaglandin-endoperoxide synthase 2	-2.14
Stom	Stomatin	3.07
Tgfb3	Transforming growth factor & beta 3	-2.04
Tnf	Tumor necrosis factor	2.89
Trim30a	Tripartite motif-containing 30A	3.69

Prioritized Genes Only for the KO-Activation

Gene Symbol	Definition	Log2(FC)
Angpt1	Angiopoietin 1	-4.96
Ccl4	Chemokine (C-C motif) ligand 4	2.06
Ccr5	Chemokine (C-C motif) receptor 5	2.16
Ceacam1	Carcinoembryonic antigen-related cell adhesion molecule 1	-2.08
Ddx58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	2.02
Il23r	Interleukin 23 receptor	6.53
Lgals9	Lectin & galactose binding & soluble 9	2.06

Mmp12	Matrix metallopeptidase 12	-5.76
Tgfb2	Transforming growth factor & beta 2	4.94
Tlr9	Toll-like receptor 9	2.06
Trim6	Tripartite motif-containing 6	4.87

6.3.2 Conclusion

Overall, the transcriptomic analysis succeeded to detect significant nuances in the level of biological processes and mechanisms, verifying the precedent findings that TRAIL modulates NK cells by repressing IFN-γ production and promoting IL-15-dependent granzyme B expression. Concerning the usage of BioInfoMiner workflow, the results of Reactome enrichment analysis confirm that the usage of strict p-value thresholds in pathway analysis could block significant biological information. Specifically, the enriched terms that verify the non-canonical regulation of PI3K/AKT signaling pathway by TRAIL, are prioritized with p-values greater that 0.05, which are considerably high values for traditional statistics to reject the null hypothesis. Thus, pathway analysis could provide useful insights (information and not noise) about the examined phenotype, even with moderate enrichments. That evidence supports the discussion in section 3.3.2, about the development of an information-based pathway analysis methodology, able to extract highly informative with minimum noise results, without the usage of statistical thresholds.

Chapter 7

Case Study 2: Protein Homeostasis Imprinting Across Evolution

This chapter presents an approach to evaluate proteostasis during evolution, using semantic analysis to classify thousands of species from different taxonomic domains. The main idea was to construct species-specific semantic profiles, related to proteostasis machinery, and compare them to assess the reliability of proteostasis as an evolutionary marker. Different tasks were performed to deconvolute the differentiated mechanistic components among Archaea, Bacteria and Eukaryotes and compare the efficiency of proteostasis with other phylogenetic criteria and molecular mechanisms. Currently, the initial manuscript of that study has been submitted in bioRχiv [202] and its final version is under revision.

7.1 Introduction

Protein homeostasis (a.k.a. proteostasis) refers to a complex and interconnected network of processes that affects both expression levels and conformational stability of proteins in cells by controlling their biogenesis, folding, trafficking and degradation within and outside the cell. The molecular mechanisms controlling proteostasis, are implicated in cell fitness, aging and contribute to disease onset. From lower Bacteria to humans, the molecular components that control proteostasis (i.e. the proteostasis network - PN) include protein synthesis, co/post-translational protein folding, quality control, degradation, as well as adaptive

signaling to proteostasis imbalance [203]. A plethora of proteins participate cooperatively in these mechanisms (ribosomal, transcription factors, kinases, chaperones, complex proteolytic machineries, etc.) to regulate the protein quality control. The PN was subjected to evolutionary pressures for each organism to cope with intrinsic and extrinsic demands. Evolution is shaped by i) the genome complexity, ii) the post-translational modifications repertoire, iii) the presence of sub-cellular compartments and iv) the emergence of multi-cellular organisms and cell differentiation. Each of these constraints increased the necessity for updated and adaptive mechanisms to ensure protein homeostasis [204].

The construction of an elucidative PN phylogenetic dendrogram needs to take into consideration the diversity of PN across species of various taxa. Traditional phylogenetic approaches use the sequence of conserved genes or proteins (or groups of them) as standard references rather than their functional identities to form ancestral lineages and identify speciation events. In theory, heat shock proteins (HSPs), that exert fundamental roles in maintaining protein homeostasis, could be considered as the appropriate markers to reflect PN evolution. Some HSP families (e.g. HSP40 and HSP70) are highly represented in most cells and the nature of this representation might reflect the underlying evolutionary relationships - e.g. 3 members of HSP40 in *E. coli* and 49 members in *Homo sapiens* [205, 206]. However, the informational content of their sequences remains unable to provide insights about functional evolution of proteostasis. Thus, a different vocabulary is necessary to exploit their functional profiles and the subsequent structured networks.

Even though studies have reported computational models of proteostasis in Bacteria (e.g. E. coli [207]) and Eukaryotes [208], an overall layout of proteostasis evolution was lacking. Herein, a novel approach was implemented to assess how the functional components of proteostasis evolved during the evolution, using semantic analysis. The main idea was to estimate the differentiation of proteostasis across multiple species, comparing the respective sets of ontological terms, using appropriate semantic analysis operators (chapter 2). However, the large complexity of PN is only partly recorded by the available ontologies. In this sense, the semantic representation and annotation of the PN is not achieved, either in terms of functional vocabulary or at the organism level. To overcome that limitation, species-pertinent gene lists, associated with the main mechanisms of proteostasis, were assembled through a supervised, multi-phase workflow. The translation of these sets into biological processes, using

BioInfoMiner, led to the required groups of semantic terms to commence the study of proteostasis evolution. Performing various computational tasks, the present study targeted to evaluate the PN as an evolutionary marker, compare its discriminative power with other criteria (ribosomal RNA sequences - rRNA and HSPs) and elucidate the underlying mechanistic differentiations among the main taxonomic super-kingdoms.

7.2 Methods

7.2.1 Data Acquisition

The selection of proteostasis-related genes for multiple species aimed to include genes strongly associated with PN key components, such as protein folding, degradation, endoplasmic reticulum, autophagy and associated signaling pathways (Fig. 7-1 A). In this way, seed gene lists were inferred for eight eukaryotic model organisms Table 7.1) and a generic list for Prokaryotes. Regarding Eukaryotes, homology mappings were used to retrieve putative, functionally similar genes, from the Ensembl repositories [128]. The model organisms were used to detect homologies with species belonging to the same generic taxonomy (e.g. *Arabidopsis thaliana* was used as reference organism for plants). Concerning the prokaryotic world, the automated search and association of homologies was facilitated by the same nomenclature system across the whole domain. Hence, the initial generic gene set was used as the base to construct the proteostasis profile for thousands of Bacteria and Archaea, exploiting the repository of UniProt Knowledgebase [209]. In parallel, another search for ribosomal RNA sequences and heat shock proteins of 40kDa and 70kDa families was performed. Ribosomal sequences (18S and 16S rRNA) were retrieved from the ENA repository [210] and heat shock proteins' data were collected from Ensembl and UniProt Knowledgebase. Organisms which met the following criteria were included in the analysis: i) quality of the genomic annotation in the GO-BP corpus; ii) availability of gene sequences of 16S (for Prokaryotes) and 18S (for Eukaryotes) rRNAs and iii) at least one annotated amino acid sequence of HSP40 (dnaJ) and HSP70 (dnaK) proteins. The final dataset comprised of 437 organisms (94 Eukaryotes, 250 Bacteria and 93 Archaea; Table C.1).

7.2.2 Pathway Analysis

The pathway analysis step of BioInfoMiner (chapter 3) was used to construct the PN profiles of species, using the Biological Process domain of Gene Ontology (GO-BP). Thus, a GO annotation collection was created for each selected organism and stored in MongoDB, to be accessible from BioInfoMiner workflow. Concerning the analysis, two distinct criteria were adopted to define the list of prioritized semantic terms. Hypergeometric p-value was determined constant at 0.05, while adjusted p-value was set to 0.05 and if an organism had fewer than one hundred terms satisfying that threshold, then the first hundred terms was selected, in order to have similar order of magnitude size of lists for all species.

Table 7.1: Eukaryotic Model Organisms used for Data Acquisition

Species	Phylum	Class	Ensembl Host
Homo sapiens	Chordata	Mammalia	ensembl.org
Danio rerio	Chordata	Actinopteri	ensembl.org
Gallus gallus	Chordata	Aves	ensembl.org
Xenopus tropicalis	Chordata	Amphibia	ensembl.org
Arabidopsis thaliana	Streptophyta	Magnoliopsida	plants.ensembl.org
Saccharomyces cerevisiae	Ascomycota	Saccharomycetes	fungi.ensembl.org
Caenorhabditis elegans	Nematoda	Chromadorea	metazoa.ensembl.org
Drosophila melanogaster	Arthropoda	Insecta	metazoa.ensembl.org

7.2.3 GO Graph Annotation and Standardization

GO has inherent inconsistencies regarding the structure and the depth of its branches [211], generating bias that hampers the comparative analysis. Some graph areas are more expanded than others, due to extensive annotation. This results in distorted, descriptive capacity, regarding the degree of specification that each term bears. Furthermore, the depth of genomic annotation differs among species. Different research communities have developed meticulous genomic annotations for model species, emphasizing on specific components of cellular physiology, according to specific characteristics of each organism [211]. On the

other hand, the vast majority of organisms has been annotated only through electronic processing [49]. This causes inconsistencies regarding the semantic network profile describing any biological process, across species. Even, taxonomically proximal species could have divergence in annotation coverage [212]. All these predicated upon a standardized version of GO, suitable for comparative analyses, balancing the depth of annotation among species. Its construction was performed setting specific thresholds for the IC (equation 2.1) and SV (2.3) values of ontological terms. GO-BP graph topology was bounded, using these two measures. The twentieth (20^{th}) percentiles of IC and SV distributions were defined as the lower bounds (Fig. 7-1 C). Terms exceeding these thresholds were trimmed and substituted with their most proximal ancestors, conforming to these rules. Then, the PN profiles were projected on the standardized version of GO, modifying its content with valid terms (Fig. 7-1 B).

7.2.4 Comparative Analysis

The implementation of phylogenetic network analysis was performed through the calculation of semantic similarities of the PN profiles. BioInfoMiner library was used for that task (section 3.2). A unified pair-wise similarity matrix was constructed for all the significant terms and afterwards the average scores were generated for each pair of organisms. In order to avoid bias of specific pair-wise measures, the similarity of two terms was calculated by averaging the scores of Rensik (equation 2.12) (based on MICA (2.5) and XGraSM (2.11) concepts) and AggregateIC (2.18). The semantic similarity of two organisms was calculated with the ABM formula (2.23). The phylogenetic dendrogram was generated using the Ward's minimum variance method [213].

7.2.5 Evaluation of Phylogenetic Analysis

Gene sequences of rRNAs were used to construct the reference phylogenetic tree, as they traditionally portray the evolutionary proximities of species [214, 215]. The ClustalW algorithm [216] was used to calculate the pair-wise distances. Furthermore, amino acid sequences of HSP40 and HSP70 were analyzed to examine their potential as surrogate evolutionary markers. HSPs of the same molecular weight could vary significantly even in the same organism. Thus, members of the same family were clustered to a consensus sequence pattern for

each organism, using CD-HIT [217] to reduce data redundancy and HMMER3 [218] to calculate the final consensus protein sequence. Then, ClustalW was used to build the respective distance matrices. Finally, the Ward's minimum variance method was used to create the phylogenies of rRNA and HSPs.

Initially, the four criteria were compared on the level of main taxonomic domains. Then, lower-level phylogenies were produced for each super-kingdom separately (Class-level for Archaea and Phylum-level for Eukaryotes and Bacteria). For each criterion, different grouping models of species, for a range of predefined number of clusters were generated, and the consistency of each model with the taxonomic classification was assessed, using the Homogeneity Score [1] [219].

7.2.6 Investigation of Proteostasis Components

Modules of BioInfoMiner library were used to cluster the GO-BP terms into generic components, according to the standardized GO graph. Similarity threshold was set to 0.175 (Resnik score), as lower values produce very generic clusters, with overly broad semantic description. A certain part of PN components was associated with the vast majority of organisms (>90%). That group was called "common components" and all the rest comprised the "different components" set. Phylogenetic analysis was performed in order to examine their contribution to the taxonomic separability of the complete PN profile.

7.2.7 Comparison of Proteostasis with Other Mechanisms

To evaluate the robustness of the proposed methodology and analyze the impact of PN onto the evolution of cellular mechanisms, another 20 conserved biological processes were analyzed. For each organism and process, a gene list was retrieved from GO-BP annotation. Pathway analysis was used to create the semantic networks and comparative analysis was performed, following the same procedure as in PN. Each phylogenetic dendrogram was divided in three clusters a priori, aiming to evaluate whether the examined mechanisms could reproduce each taxonomic domain. Their efficiency was quantified with Homogeneity [219] and Silhouette[2] [220] Scores. As a last step, PN components were excluded from

[1]The Homogeneity Score evaluates the purity of the output of an unsupervised clustering, given the reference classification of the clustered samples (their true labels).

[2]The Silhouette Score quantifies the coherence of the clusters derived from an unsupervised approach, regardless the accuracy of the final output.

these semantic profiles and the comparative analysis was performed again, in order to measure their merit as taxonomic classifiers.

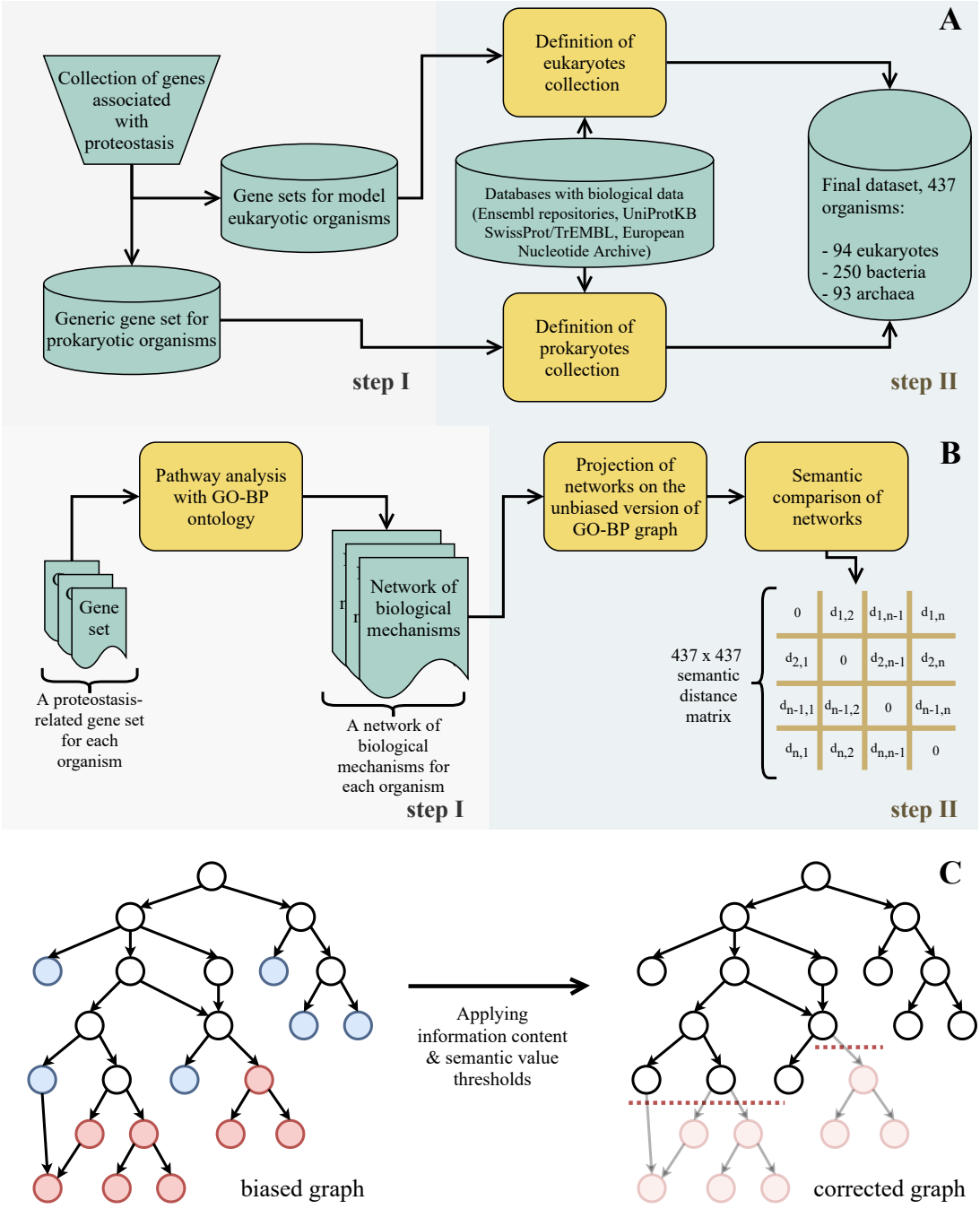


Figure 7-1: (A) Data acquisition: Proteostasis-related gene lists were defined manually for model organisms and databases were used to collect homologies and data for rRNA, and HSPs to expand organisms' collection. (B) Analysis workflow: Gene lists were translated into GO-BP networks. Semantic analysis was performed to calculate their semantic distances. (C) Construction of the unbiased GO-BP graph: Leaf terms (red and blue nodes) located in elongated branches (only red nodes) were filtered out to neutralize the knowledge representation on the graph.

7.3 Results & Discussion

7.3.1 Ribosomal RNA, HSP40, HSP70 & Proteostasis-Based Phylogenies

PN succeeded to separate the three main taxonomic domains almost infallibly (Fig. C-1), performing as accurately as rRNA sequences do (Fig. 7-2 A). PN evolution appeared less constrained than that of rRNA, which led to distantly separated super-kingdoms. This probably reflects the heterogeneity of the PN content, bisected into domain-specific components and others that are essentially conserved across evolution. Nevertheless, the pair-wise distances among species for rRNA and PN showed strong correlation (Fig. 7-2 B). Concerning the HSP-based classification, a poor correlation of the HSPs sequences with evolution was observed, as they succeeded to separate only eukaryotic and prokaryotic kingdoms, even so, not flawlessly. These findings primarily indicate that the proposed method can disclose evolutionary differences, among species of different taxonomic domain. They also demonstrate the utility of PN as a reliable evolutionary marker, able to classify species according to their main taxonomy, contrary to the limitations of HSP sequence-based approaches.

The comparison in lower-level taxonomies corroborated the findings inferred from the rRNA sequences and revealed an overall homogeneity of PN profiles in Bacteria and Archaea Fig. C-2. None of the criteria succeeded to produce a number of clusters equal to that of the reference taxonomic groups, verifying that species of different lower-level taxonomies share similar profiles. rRNA sequences represented the most effective measure. The PN-based classification indicated that Bacteria and Archaea share PN components, as the produced models of clusters have constantly low Homogeneity Score. The eukaryotic PN was more complex, encompassing variations that coherently segregated the species, given their Phylum. HSP-based clusters were similar to those obtained with rRNA sequences in Prokaryotes, but declined among Eukaryotes, probably due to the high variation of protein families possessing 'species-specific' profiles. At last, PN-based species organization provided a much more robust classification at the Phylum level in Eukaryotes, compared to HSPs.

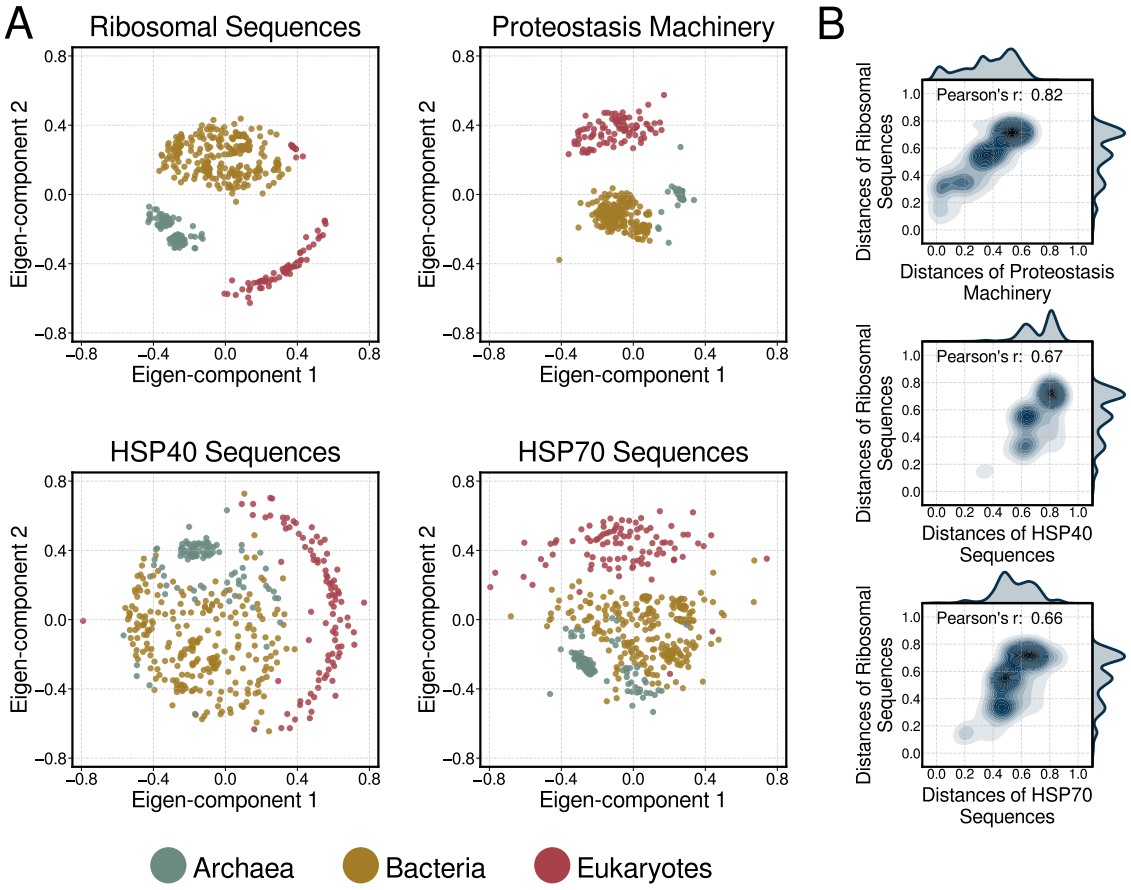


Figure 7-2: Comparison of rRNA, HSP40, HSP70 and proteostasis-based phylogenies. (A) Two-dimensional representation of organisms in function with their taxonomy. The derived evolutionary distance matrix of each criterion, was transformed into a 2-dimensional space through the Multidimensional Scaling (MDS) algorithm [221], reflecting the two larger, dimensions of observed variation (termed eigen-components). Similarity distances of each organism from the centroid of the single class problem are projected in those exploratory scatter plots. (B) Pearson correlation of pair-wise distances of rRNA sequences with the other three measures. Correlation is unbiased from taxonomic domain sizes, as 80 randomly selected species from each domain were used for the calculation.

7.3.2 Tracing Evolution Based on Proteostasis Components

The semantic clustering of proteostasis-related GO-BP terms, indicated that the major PN components could be classified into 56 distinct generic groups (Fig. 7-3). The output list revealed that many, if not all cellular processes, are connected to the PN, especially for Eukaryotes. The conserved PN component across the vast majority of species is linked with protein production and folding, responses to external or internal stimuli, activation or repression of anabolic and catabolic

processes to maintain cell homeostasis and the localization of macromolecules. In addition, specific regulatory functions were enriched in Eukaryotes, such as proteins associated with programmed cell death or signaling pathways. Enrichment in the response to endoplasmic reticulum (ER) homeostasis imbalance in Eukaryotes was effective in all eukaryotic species tested, pointing out this compartment as a hotspot for proteostasis. Similarly, specific enrichment of membranous organelles associated mechanisms, such as mitochondrial organization or autophagy, were exclusively identified in Eukaryotes, whereas metabolic pathways such as nitrogen compounds processes, amide metabolism or even protein unfolding were associated to Prokaryotes. Using the segregation of PN components in "common" and "different" sets, two-dimensional representations of species were generated for the whole PN and the two sub-groups of PN components Fig. 7-4. The whole PN functional profile produced independent clusters for each taxonomic domain. As expected, these clusters exhibited an increased density when analyzing the common PN and showed further expansion when analyzing the differential PN.

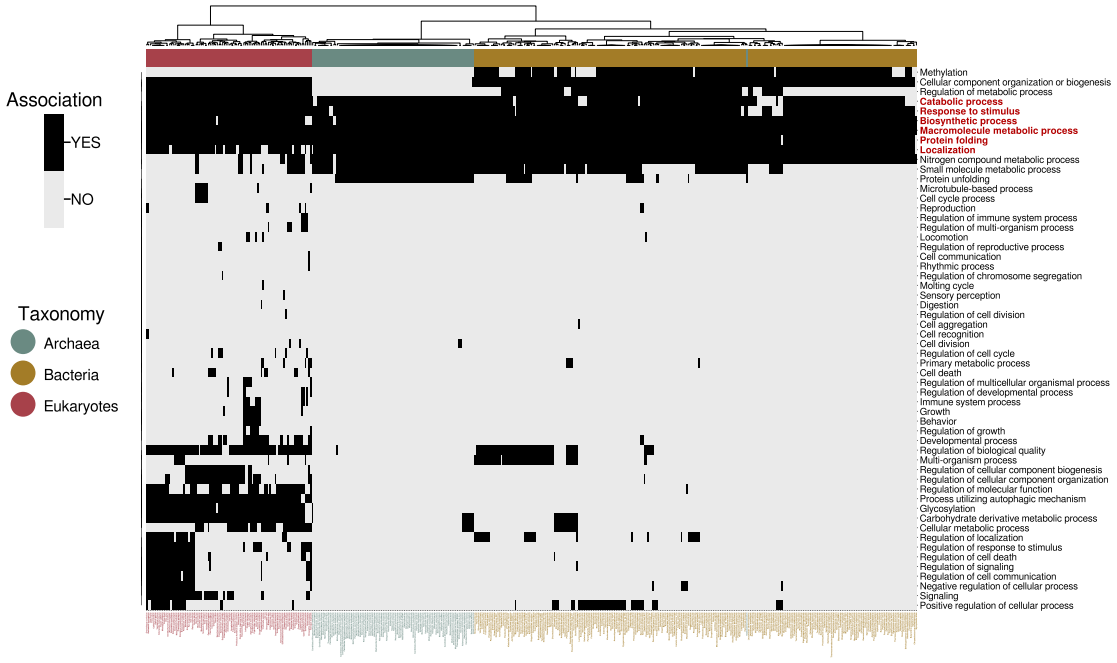


Figure 7-3: Association matrix of proteostasis' main components with organisms' collection. These components were revealed performing a semantic clustering of enriched pathways on the GO-BO graph. Each pathway was aggregated to more generic terms, ending up to that set of systemic processes. The red section refers to the mechanisms associated with more than 90% of species, named "common components".

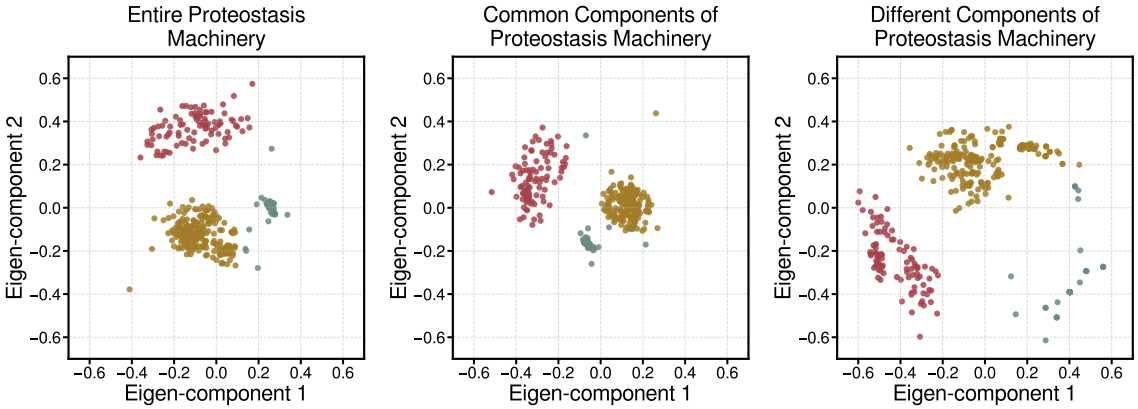


Figure 7-4: Two-dimensional representation of organisms based on proteostasis profiles, using the Multidimensional Scaling (MDS) algorithm [221]. "Common" and "different" mechanistic components were examined to investigate their contribution to the separation of taxonomic domains.

7.3.3 Impact of Proteostasis Evolution on the Nature of Other Mechanisms

The semantic analysis of the other biological processes showed that these which potentially constitute one of the mainspring features of species evolution or even those which have been differentiated by adopting more intricate functional networks due to the evolutionary pressure, are adequately informative to separate the three super-kingdoms. Their list include "tRNA processing", "cell compartmentalization", "lipid metabolism", "methylation" and regulatory networks. Some performed marginally better in terms of accuracy, as taxonomic metric, compared to proteostasis or the rRNA sequences. Nevertheless, their phylogenies exhibited lower Silhouette Score values comparing to PN, which means that the produced clusters are sparser (especially within Prokaryotes), having lower coherence and, consequently, the phylogenetic trees contain broader clades. On the other hand, parts of the aforementioned processes or those with narrower functional networks showed weaker performance, concerning the clustering efficiency, especially due to their strong commonalities among the Prokaryotes. Homogeneity Score measurements were below 0.8, principally because large groups of Archaea were classified in Bacteria, and vice versa. The removal of PN components from their profiles revealed the instrumental role of proteostasis as a powerful indicator of the cellular and organismal adaptations to evolutionary cues. The taxonomic performance of the majority of processes was conspicuously decreased.

Only the regulatory processes and cellular component assembly retained adequate information to distinguish accurately the taxonomic domains, implying that their mechanistic framework can differentiate the various taxonomies. In general, all the processes with accurate performance suffered from low Silhouette Scores, and some lost their phylogenetic congruity. Improvements of Homogeneity and Silhouette Score values for some biological processes in the defective group were meaningless, as they did not address the phylogenetic comparison. This analysis indicated that the PN contributes to various, interconnected pathways, possessing pivotal roles in cell homeostasis and functions.

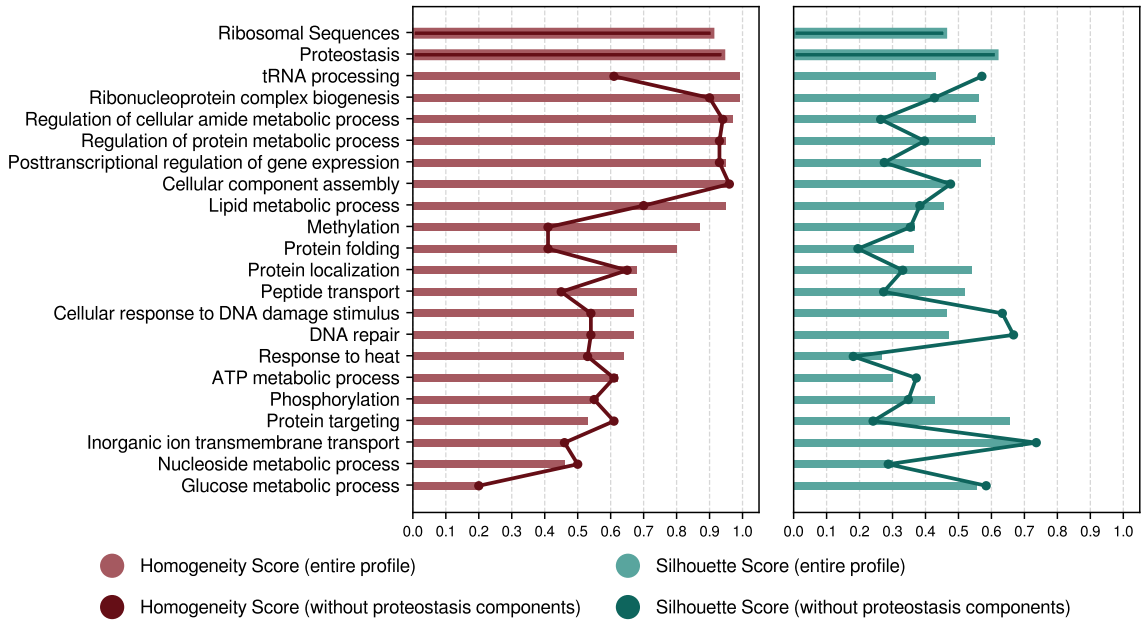


Figure 7-5: Proteostasis contribution to other evolutionary conserved mechanisms. Homogeneity score (red bars) refers to the separating ability of each biological process regarding the three main taxonomic domains, through the respective semantic network, derived from the pathway analysis of related genes. Silhouette score (green bars) indicates the degree of cohesion of cluster inference, by measuring the trade-off between intra- and inter-distances of each cluster member. Bars display these scores for the entire machinery of each process whereas the solid lines illustrates the same scores calculated after the removal of proteostasis related components from the gene machinery of each process.

7.3.4 Conclusion

The usage of BioInfoMiner library facilitated the implementation of various analytical steps in this study. Its integrated modules enabled the automated

construction of hundreds species-specific GO-BP $Graph$ instances, the standardization of GO annotation and the creation of phylogenetic trees based on the semantic analysis of GO-BP for proteostasis and another 20 highly-conserved biological processes. Such a versatility was the reason that the core algorithms of BioInfoMiner were designed from the ground up and their individual functional components were constructed in an object-oriented logic.

The results shown that PN semantic networks provide reliable information about the evolution of main taxonomic domains, performing as accurately as rRNA sequences and better than isolated PN components, such as the HSPs. The efficiency of the PN metric dropped in Bacteria and Archaea, likely due to their poor annotations compared to the Eukaryotes. A way to solve that issue could be the massive informational integration from vocabularies with better functional annotations, such as the MetaCyc database [44]. The deconvolution of PN components revealed that many, if not all cellular processes, are connected to the proteostasis, especially for Eukaryotes. Also, the role of proteostasis as a robust indicator of cellular and organismal adaptations to evolutionary cues, is highlighted by the taxonomic under-performance that the other mechanisms linked to PN exhibit when the PN components are excluded. As such, PN encompasses all the biological information to categorize species according to their complexity, and acts as a fingerprint of evolution.

Overall, this study proposed a novel approach for phylogenetic comparison, which uses the semantic graph as a new metric in order to evaluate the complexity of protein homeostasis. This stands as a novel strategy for taxonomic classification, which assess divergence of the topological similarity of ontological trees, providing species-specific functional profiles rather than analyzing individual sequences. Thus, the semantic interpretation of gene and protein sets for different species could provide biological insights about the impacts of the evolutionary pressure and the extent of conservation of mechanisms among different taxonomies.

Chapter 8

Case Study 3: Semantic Interpretation of the SARS-CoV-2 Interactome

The spread of SARS-CoV-2 across the world evolved in an unprecedented enemy for the public health. Myriad scientific studies started immediately to describe its cycle of life and the underlying interactions with human cells, targeting to uncover potential therapeutic targets. One of the most seminal studies during the first year of COVID-19 pandemic was that of Gordon et al. [222], about the SARS-CoV-2-human protein-protein interactions network. The study predicted with high confidence that 332 human proteins interact physically with viral proteins in different stages of viral cycle life. This chapter concerns the semantic interpretation of that protein set (in short SARS-CoV-2 interactome), as a part of a review study about the interaction of SARS-CoV-2 with different components of secretory pathway. The bioinformatic analysis was comprised of the interpretation of the SARS-CoV-2 interactome with BioInfoMiner and its detailed comparison with another 11 viral interactomes, using different biomedical ontologies. The project was elaborated in terms of a five-month secondment in the Centre in the Fight Against Cancer (Centre de Lutte Contre le Cancer; CLCC) of Center Eugène Marquis in Rennes (France) and the results have been published in [223].

8.1 Introduction

On March 11, 2020, the World Health Organization (WHO) declared COVID-19, an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), as a pandemic. After one year, over 117 million infection cases and over 2.6 million confirmed deaths have been reported worldwide [224, 225], while SARS-CoV-2 is still (March 2021) an ongoing serious threat for global health. The initial clinical features showed that SARS-CoV-2 affects the respiratory system and the common mild symptoms include fever, dry cough and fatigue. Severe cases are manifested mainly in older men with co-morbidities [225] and patients develop pneumonia, pulmonary oedema, acute respiratory distress syndrome and multiple organ failure, which could lead to death [226]. Historically, six other coronaviruses have been recorded to cause human infectious diseases [227]. Apart from those causing mild symptoms, SARS-CoV (a.k.a. SARS-CoV-1) and MERS-CoV (Middle East respiratory syndrome-related coronavirus) were responsible for major outbreaks and hundreds of deaths in specific territories of Asia during 2002-2003 and 2012-2015 respectively.

SARS coronaviruses and MERS-CoV belong to the genus of *Betacoronaviruses* and their genome is a single-stranded, positive-sense RNA comprised of 26 to 32 kb [227]. Phylogenetic analysis of SARS-CoV-2 samples, isolated in the place where the outbreak started in December 2019 (Wuhan, Hubei province, China) revealed that the genome sequence similarity between SARS coronaviruses is about 79% and that of SARS-CoV-2 with MERS-CoV around 50% [228]. SARS viruses seem to have approximately identical genomes, explaining their common classification in the A lineage of *Betacoronaviruses*, while MERS-CoV belongs in the C lineage. Even so, their genomes have the general organisation of 5′-leader-UTR-replicase-S(Spike)-E(Envelope)-M(Membrane)-N(Nucleocapsid)-3′UTR-poly(A)tail [229, 230]. The 5'-end open reading frame (ORF; ORF1a/b complex) extends to the two-thirds of the whole genome and encodes polyproteins (pp1a and pp1ab), which are further cleaved into a variety (1˜6) of non structural proteins (nsps), that are necessary for the processes of transcription and replication. The following ORFs, located towards the 3′-end, produce the structural proteins of spike (S), nucleocapside (N), membrane (M), and envelope (E), as well accessory factors, which have pivotal role in the viral pathogenesis, since they aid the assembly and transport of new particles.

Like the SARS-CoV, the infectious cycle of SARS-CoV-2 initiates with the

attachment of its spike (S) protein on the target cell surface [231, 232]. S protein consists of two subunits (S1 and S2). S1 contains a receptor-binding domain (RBD) which recognizes the angiotensin-converting enzyme 2 (ACE2) as its cellular receptor and binds on it. Then host proteases (such as transmembrane protease serine 2; TMPRSS2 [233], lysosomal proteases cathepsins B and L [234] and extracellular furin-like enzymes [235]) cleave the spike protein at the S1/S2 boundary. That proteolytic reaction changes the conformation of S2 unit, which subsequently causes the fusion of viral envelope with host cell membrane and the entry of viral genome in the cell. As SARS-CoV-2 RNA is positive-strand, it serves as mRNA for the 5'-end ORF and hijacks the host-cell ribosomes to encode the two poly-protein molecules. These poly-peptides are proteolytically processed into the individual nsps via host and viral proteases (papain-like protease; nsp3 and 3C-like protease domain; nsp5) [236]. The non structural proteins form the replicase/transcriptase complex (RTC), which is the crucial factor for virus reproduction. RTC contains the RNA-dependent RNA polymerase (RdRp; nsp12), which catalyzes the replication of viral genome [237], while other embedded enzymes are responsible for the production of structural and accessory proteins, using as transcription template a group of sub-genomic RNAs (sgRNAs) at the 3'-end of the viral genome. The produced S, M, and E proteins undergo diverse post-translational modifications in Endoplasmic Reticulum (ER) and then they are transported to ER–Golgi intermediate compartment (ERGIC). The N proteins encapsidate the new genomic RNA and that complex is assembled with the structural proteins to form the virion in ERGIC [238]. Finally, mature virions are proceed through the Golgi complex and later stations of the secretory pathway to the extracellular space.

Although the main aspects of SARS-CoV-2 life cycle have been comprehended, hitherto there is not any effective remedy to curb the spread of pandemic. Potential therapies could be separated in two main categories, targeting either human or viral proteins [239]. Concerning the first category, a strategy with great potential to reveal drug targets is to study the virus-host protein-protein interactions (PPIs) network. Gordon et. al [222] performed such an analysis to uncover the biomolecules and mechanisms of human cells, which interact physically with the viral proteins and contribute to the regulation of infection. Finally, they identified a high-confidence SARS-CoV-2-human PPIs network, consisting of 26 viral and 332 human proteins (henceforth referred as SARS-CoV-2 interactome). The authors used the HEK293T cell line as model and

taking advantage of affinity purification mass spectrometry (AP-MS) assays, they detected 66 virus-host interactions, targeted by 69 existing FDA-approved drugs. In this work, BioInfoMiner was used to provide a systemic overview of the SARS-CoV-2 interactome with various ontologies and interrogate the centrality of human proteins on the consensus semantic networks. Furthermore, a comparative analysis of various interactomes, related to pathogenic viruses, was performed, in order to estimate their functional relevance with that of SARS-CoV-2.

8.2 Methods

8.2.1 Data Acquisition

The SARS-CoV-2 interactome, as well those of other well-known pathogenic viruses, were retrieved from recent studies of Krogan laboratory. The researchers have determined a number of high-confidence virus-human PPI networks, targeting to interpret how viruses co-opt human machinery during infection, uncover causal functions and therapeutic targets. Their experimental protocols are based on the implementation of affinity purification-mass spectrometry (AP-MS) approach [240], using the viral proteins as baits to identify human proteins that uniquely interact with them. The quantification of PPIs from the extracted proteomic datasets has been performed with various methodologies, such as MiST [241], CompPASS [242] and SAINTexpress [243]. As a whole, virus-human PPI networks have been identified for Human Immunodeficiency Virus (HIV) [241], Hepatitis C Virus (HCV; separate experiments for Huh7 and HEK293T cell lines) [244], Herpesvirus [245], Papillomavirus [246], Ebola [247], Dengue and Zika [248], West Nile Virus (WNV) [249], Enteroviruses (such as Coxsackievirus A10 and Rhinovirus C15) [250] and most recently SARS-CoV-2 [222]. All the interactomes were retrieved from the respective supplementary files.

8.2.2 Semantic Interpretation and Comparative Analysis

BioInfoMiner was used to perform the semantic interpretation of human proteins of SARS-CoV-2 interactome, with three distinct ontological vocabularies (GO-BP, Reactome and MGIMP), and prioritize them on the produced semantic networks. Additionally, pathway analysis was performed for all the aforemen-

tioned interactomes, to measure their enrichment of the vertices of ontological graphs. The comparative analysis was performed similarly to that of PN profiles in chapter 7, without the construction of standardised ontological trees, as all the protein lists referred to the proteome of *Homo sapiens*. This task was executed separately for each ontology, enabling the comparison of interactomes under different semantic perspectives. A unified pair-wise similarity matrix was constructed for all the over-represented terms using the average score of their Rensik (equation 2.12) (based on MICA (2.5) and XGraSM (2.11) concepts) and AggregateIC (2.18) similarities. Then, the semantic similarity of each pair of viral-host interactomes was calculated with the ABM formula (2.23) and the phylogenetic dendrogram was generated, using agglomerative clustering with Ward's minimum variance method [213].

8.3 Results & Discussion

8.3.1 The Semantic Landscape of SARS-CoV-2 Interactome

The results of pathway analysis of SARS-CoV-2 interactome with GO-BP, Reactome and MGIMP ontologies are presented in Table D.1, D.2 and D.3. Concerning the functional interpretation, GO-BP enriched graph indicates that SARS-CoV-2 physically interacts with multiple nodal mechanisms of cellular functionality. Particularly, enriched elongated branches are associated with cellular components organization, regulation of catabolic process and RNA stability, membrane docking, regulation of intracellular transport, protein folding, localization, transport and metabolism and response to ER stress. Overall, these processes are essential for the properly regulated cell cycle and cellular homeostasis. Reactome terms verify the involvement of protein localization and post-translational modification, metabolic process of RNAs and cell cycle. Evidently, SARS-CoV-2 interact with a wide range of host factors to succeed its reproduction, harming cellular physiology without interacting directly with immune system molecules (such as cytokines). Enriched terms of MGIMP belong to various deficiencies of circulatory system, neurodegeneration and morphological abnormalities of cellular components, indicating that many SARS-CoV-2 interactors have been associated with abnormal phenotypic traits. Of note, early studies during the first months of pandemic associated COVID-19 with potential development of cardiovascular diseases [251, 252], underlying their

high prevalence among the patients. Interestingly, recent studies alleged a strong association between COVID-19 and neurological symptoms, implying that the neuroinvasive potential of SARS-CoV-2 could affect the brain and contribute to the development of neurodegenerative and neuropsychiatric diseases [253–255].

8.3.2 Semantic-Based Prioritized Proteins of Infected Human Cells

Concentrating the lists of prioritized proteins for the three ontological vocabularies (Table 8.1), 18 proteins were found as nodal in two or more semantic networks. RABA8 and RAB1A, both members of Ras oncogene family, were found prioritized in all three and two ontologies respectively. Many other members of Rab family are included in SARS-CoV-2 interactome (RAB10, RAB14, RAB18, RAB2A, RAB5C, RAB7A), and interact with nsp7. Also Rab protein signal transduction has been revealed as a significant pathway (based on GO-BP), underpinning the statement that those small GTPases have important role in SARS-CoV-2 infection. Rab proteins are essential factors of secretory pathway, as they control membrane trafficking [256] and they have been proposed as a promising targets for the development of antiviral therapies [257]. Another important protein family with many highly prioritized members is that of nucleoporins (NUP54, NUP58, NUP62, NUP88, NUP98, NU210, NUP214, RAE1). In general, nucleoporins control the nucleocytoplasmic transport, forming the nuclear pore complex (NPC) [258]. Although SARS-CoV-2 does not enter in nucleus to proliferate, it interacts with NPC to promote its replication. Its accessory protein orf6 localizes at NPC and directly interacts with NUP98-RAE1, impeding the nucleocytoplasmic transport of various macromolecules. Recent study found that orf6 cause the accumulation of mRNA transporters in nucleus and subsequently hampers the proper cellular functionality [259], while another uncovered the blockage of STAT molecules, which induce the transcription of IFN-stimulated genes (ISGs), related to immune response [260]. The family of integrins (such as ITGB1) has been reported as alternative cell receptors of SARS-CoV-2 spike protein and thus they constitute another candidate therapeutic target [261, 262]. The prioritization of proteins related to ER (ERLEC1, OS9, WFS1) and Golgi apparatus (GOLGA2 and GORASP1), and the existence of enriched terms, such as ”ERAD pathway”, ”response to endoplasmic reticulum stress” and ”Golgi organization” imply that the non structural proteins of SARS-CoV-2 may con-

trol ER–Golgi homeostasis by acting on diverse related components [223, 263]. Another highly prioritized gene is HMOX1, whose encoded enzyme catalyzes the first step of the oxidative degradation of the heme group. Studies have mentioned the anti-inflammatory role of HMOX1 due to its up-regulation under conditions of oxidative stress [264], whereas its polymorphism has been marked as a potential genetic factor for susceptibility to COVID-19 [265]. HMOX1 binds with SARS-COV-2 orf3 protein and recently it was proposed as a drug target for the prevention and treatment of COVID-19 [266]. Also, the list of prioritized proteins includes some kinases (PRKACA, PRKAR2B, MARK2, BCKDK, NEK9), which are indispensable factors for the viral infection, as they phosphorylate viral proteins, contributing to the process of replication [267]. Numerous of kinases are targets of FDA-approved drugs and new studies suggest the repurposing of their inhibitors to construct therapeutic solutions for COVID-19 [268, 269].

Table 8.1: Prioritized Proteins of the SARS-CoV-2 Interactome

	Gene Symbol	Definition	Ontologies	L1000 Perturbed
1	RAB8A	RAB8A, member RAS oncogene family	GO-BP, Reactome, MGIMP	✓
2	HMOX1	Heme oxygenase 1	GO-BP, MGIMP	✓
3	NUP98	Nucleoporin 98	GO-BP, Reactome	✓
4	NUP210	Nucleoporin 210	GO-BP, Reactome	✓
5	ITGB1	Integrin subunit beta 1	GO-BP, MGIMP	✓
6	RAB1A	RAB1A, member RAS oncogene family	GO-BP, Reactome	
7	NUP214	Nucleoporin 214	GO-BP, Reactome	✓
8	GOLGA2	Golgin A2	GO-BP, Reactome	✓
9	RAE1	Ribonucleic acid export 1	GO-BP, Reactome	✓
10	NUP62	Nucleoporin 62	GO-BP, Reactome	✓
11	IDE	Insulin degrading enzyme	GO-BP, MGIMP	✓
12	NUP54	Nucleoporin 54	GO-BP, Reactome	
13	NUP58	Nucleoporin 58	GO-BP, Reactome	
14	NUP88	Nucleoporin 88	GO-BP, Reactome	✓
15	RBX1	Ring-box 1	GO-BP, Reactome	
16	CDK5RAP2	CDK5 regulatory subunit associated protein 2	Reactome, MGIMP	
17	PRKACA	Protein kinase cAMP-activated catalytic subunit alpha	GO-BP, Reactome	✓
18	GORASP1	Golgi reassembly stacking protein 1	GO-BP, Reactome	
19	AKAP9	A-kinase anchoring protein 9	Reactome	✓

Continued on next page

Table 8.1 – continued from previous page

	Gene Symbol	Definition	Ontologies	L1000 Perturbed
20	SLC25A21	Solute carrier family 25 member 21	MGIMP	
21	RALA	RAS like proto-oncogene A	MGIMP	
22	TUBGCP3	Tubulin gamma complex associated protein 3	Reactome	
23	TUBGCP2	Tubulin gamma complex associated protein 2	Reactome	
24	ERLEC1	Endoplasmic reticulum lectin 1	GO-BP	
25	CEP250	Centrosomal protein 250	Reactome	
26	GNB1	G protein subunit beta 1	MGIMP	✓
27	RDX	Radixin	MGIMP	✓
28	PLAT	Plasminogen activator, tissue type	MGIMP	✓
29	EIF4H	Eukaryotic translation initiation factor 4H	MGIMP	
30	FAR2	Fatty acyl-CoA reductase 2	MGIMP	✓
31	VPS11	VPS11, CORVET/HOPS core subunit	GO-BP	
32	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif 1	MGIMP	✓
33	RHOA	Ras homolog family member A	GO-BP	✓
34	FBN1	Fibrillin 1	MGIMP	✓
35	CNTRL	Centriolin	Reactome	✓
36	BRD2	Bromodomain containing 2	MGIMP	✓
37	RAB7A	RAB7A, member RAS oncogene family	GO-BP	
38	SBNO1	Strawberry notch homolog 1	MGIMP	
39	WFS1	Wolframin ER transmembrane glycoprotein	MGIMP	✓
40	NIN	Ninein	GO-BP	
41	PRKAR2B	Protein kinase cAMP-dependent type II regulatory subunit beta	Reactome	✓
42	CEP135	entrosomal protein 135	Reactome	✓
43	SCARB1	Scavenger receptor class B member 1	GO-BP	✓
44	MARK2	Microtubule affinity regulating kinase 2	MGIMP	
45	GDF15	Growth differentiation factor 15	MGIMP	✓
46	OS9	OS9, endoplasmic reticulum lectin	GO-BP	✓
47	CLCC1	Chloride channel CLIC like 1	MGIMP	
48	EXOSC3	Exosome component 3	GO-BP	
49	EXOSC2	Exosome component 2	GO-BP	
50	BCKDK	Branched chain ketoacid dehydrogenase kinase	MGIMP	✓
51	UPF1	UPF1, RNA helicase and ATPase	GO-BP	
52	CHMP2A	Charged multivesicular body protein 2A	GO-BP	

Continued on next page

Table 8.1 – continued from previous page

	Gene Symbol	Definition	Ontologies	L1000 Perturbed
53	RAB2A	RAB2A, member RAS oncogene family	Reactome	✓
54	ALG8	ALG8, alpha-1,3-glucosyltransferase	MGIMP	✓
55	RAB18	RAB18, member RAS oncogene family	MGIMP	
56	PCSK6	Proprotein convertase subtilisin/kexin type 6	MGIMP	✓
57	BAG5	BCL2 associated athanogene 5	GO-BP	✓
58	FKBP10	FKBP prolyl isomerase 10	MGIMP	✓
59	PSMD8	Proteasome 26S subunit, non-ATPase 8	Reactome	
60	HDAC2	Histone deacetylase 2	MGIMP	
61	POLA1	DNA polymerase alpha 1, catalytic subunit	GO-BP	
62	LOX	Lysyl oxidase	MGIMP	✓
63	FBLN5	Fibulin 5	MGIMP	✓
64	NINL	Ninein like	Reactome	
65	NEK9	NIMA related kinase 9	Reactome	
66	RAB10	RAB10, member RAS oncogene family	GO-BP	
67	PITRM1	Pitrilysin metallopeptidase 1	MGIMP	
68	TIMM10	Translocase of inner mitochondrial membrane 10	GO-BP	
69	GCC2	GRIP and coiled-coil domain containing 2	GO-BP	✓
70	TOR1A	Torsin family 1 member A	GO-BP	✓
71	AKAP8L	A-kinase anchoring protein 8 like	GO-BP	✓
72	SMOC1	SPARC related modular calcium binding 1	MGIMP	
73	AP3B1	Adaptor related protein complex 3 subunit beta 1	MGIMP	
74	MIB1	Mindbomb E3 ubiquitin protein ligase 1	MGIMP	
75	ATP6AP1	ATPase H+ transporting accessory protein 1	GO-BP	
76	PPT1	Palmitoyl-protein thioesterase 1	MGIMP	
77	NPC2	NPC intracellular cholesterol transporter 2	MGIMP	✓
78	TRIM59	Tripartite motif containing 59	MGIMP	
79	GIGYF2	GRB10 interacting GYF protein 2	MGIMP	
80	PCNT	Pericentrin	Reactome	

8.3.3 Semantic Clustering of Viral-Human Interactomes

This network-aided comparative analysis produced three different phylogenies for the aforementioned viral interactomes (Fig. 8-1, 8-2 and 8-3). Measuring the enrichment of the vertices on each graph, a distance matrix was calculated and the agglomerative clustering brought nearby the interactomes with semantic commonalities. The analysis shown that SARS-CoV-2 share the highest similarities with the enteroviruses (Coxsackievirus A10 and Rhinovirus C15). Particularly, two robust groups of viruses have been disclosed, whose members are consistently clustered together, in accordance with their taxonomic classification. SARS-CoV-2 and enteroviruses belong to the class of *Pisoniviricetes*. The other group consists of Dengue, Zika and West Nile Virus, whose taxonomic relevance is stronger, as all of them are included in the genus of *Flavivirus*. While Hepatitis C belongs to the same taxonomic family of the latter group, it is mainly clustered with Papillomavirus, in spite of their moderate similarity scores. The other viruses are clustered with a non-systematic consistency.

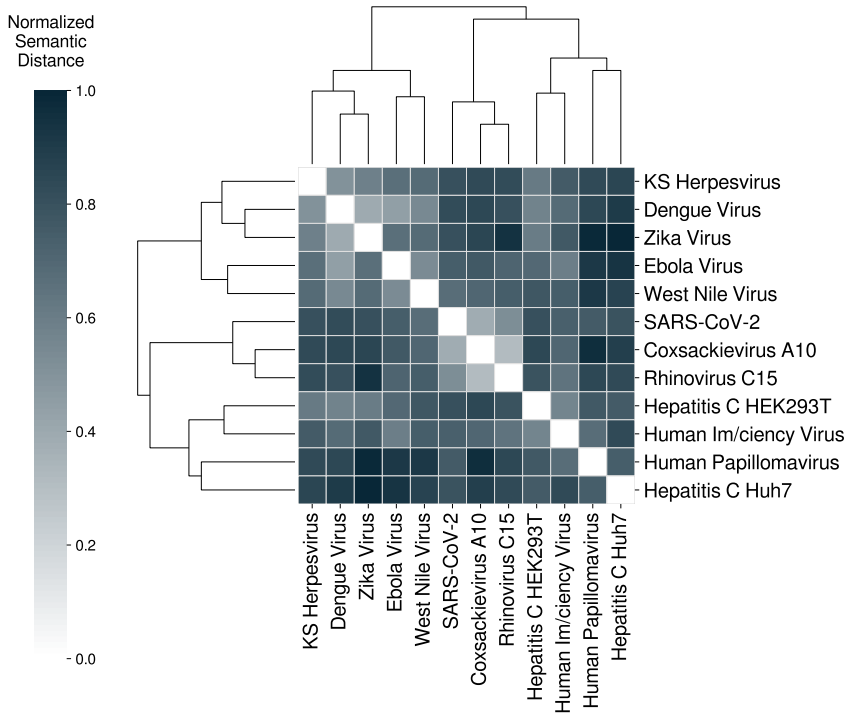


Figure 8-1: Comparative analysis of 12 viral pathogens infection models using the GO-BP vocabulary. The phylogenetic tree was constructed with agglomerative clustering, based on the semantic similarities of viral-human interactomes.

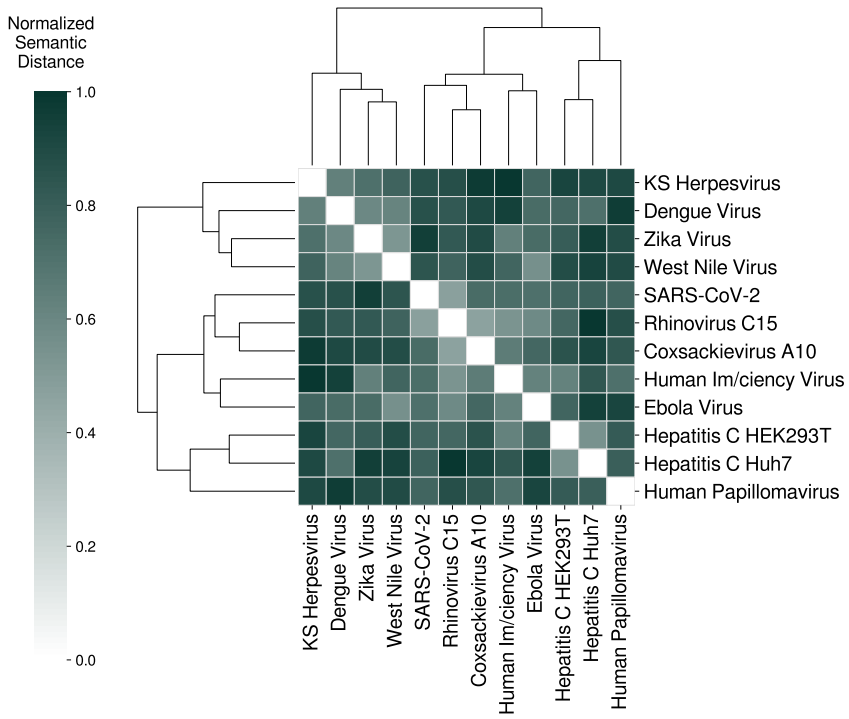


Figure 8-2: Comparative analysis of 12 viral pathogens infection models using the Reactome database. The phylogenetic tree was constructed with agglomerative clustering, based on the semantic similarities of viral-human interactomes.

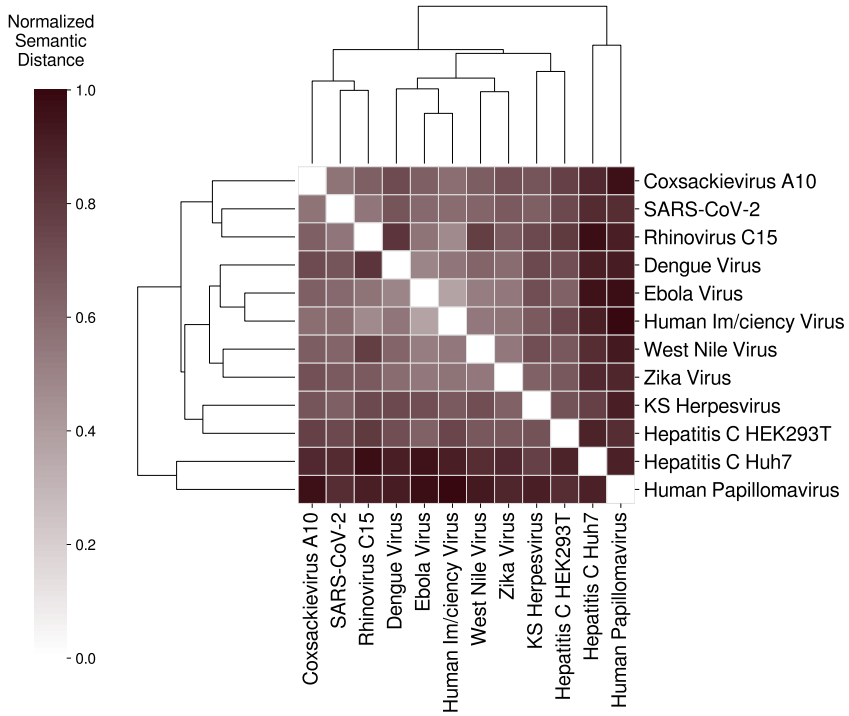


Figure 8-3: Comparative analysis of 12 viral pathogens infection models using the MGIMP ontology. The phylogenetic tree was constructed with agglomerative clustering, based on the semantic similarities of viral-human interactomes.

8.3.4 Conclusion

The interpretation of the SARS-CoV-2 interactome with BioInfoMiner constructed an integrative semantic description, using different aspects of biological knowledge. Also, the results of comparative analysis signify a correlation of SARS-CoV-2 functionalities with enteroviruses. The proteins of SARS-COV-2 interact with many pathways and mechanisms of protein homeostasis and affect mainly the intracellular transport of macromolecules, RNA stability and organelle organization. BioInfoMiner succeeded to reveal phenotypic traits caused by SARS-CoV-2 or correlated diseases, in accordance with recent reviews and experimental studies. The implementation of gene prioritization step on the three different ontologies, reduced the initial set of 332 human-virus interactors approximately four times, resulting in 80 proteins with high centrality on the respective semantic networks. The half of them have been characterized as perturbed proteins in the L1000 Connectivity Map repository [270] (Table 8.1), indicating their regulatory role in cellular functionality. A custom operator was implemented to exploit the whole vector of the derived protein signature, in order to prioritize perturbagens that affect a subset of them, ordered by the significance of their overlap with the proposed targeted gene sets of the L1000. This amount of perturbed proteins denotes a great repertoire of potential drug targets and intensifies the necessity of further pharmacogenomic and drug-repurposing studies to identify candidate drugs or drug combinations to target SARS-CoV-2.

Chapter 9

Conclusions

9.1 Discussion

The methodologies developed for the biological interpretation and comparative analysis of omics data in this doctoral thesis, are based on the idea of semantic network analysis. They utilize the semantic network in order to either draw critical conclusions about the collaborative relationships of biomolecules and pinpoint their semantic centrality or to estimate their relative distance based on a given ontological annotation. Therefore, they are applicable for any hierarchical vocabulary and any class of biomolecules, given that there is a determined mapping between them. That is to say, the proposed methodologies are exclusively data-driven and agnostic and they could be implemented for different types of omics data, regardless of the experimental values (p-values, fold changes etc).

The importance of this computational approach is stemming from two main reasons, related to the developments in the Biomedical Sciences over the last two decades. Both the evolution of omics technologies and the organization of ontologies are at a significant level of maturity, being important pillars of research in the field of Systems Biology. The high-throughput technologies are in an ongoing advancement, as novel, more detailed experimental methods are being designed and implemented, reducing the cost and time of experiments. However, the fundamental idea of omics, which is the large scale screening of biomolecules and the quantification of their expression, regulation and modification in cells, has been established as the principal methodological approach to investigate complex biological systems and phenotypes. Even the recent trend of multi-omics

studies is based on the screening of different -omes, with various experiments for the same biological system, and the combined interpretation of their data. The paradigm shift from the traditional Molecular Biology to Systems Biology has already been adopted by the industry in many applications for drug target discovery, personalized medicine and the discovery [271] of biotechnological products [272]. By the same token, biomedical ontologies have become one of the most basic means of organizing existing knowledge in various fields of Biology. The scientific community has developed and is constantly evolving appropriate languages [40, 41] for the construction of ontologies and methodologies for the automated characterization of new organisms, mainly from the kingdom of Bacteria [49]. As a result, the proposed methodologies are not related to a state-of-the-art approach in biomedical studies, but to predominantly technologies of how complex biological systems are studied and how the existing knowledge is organized to be used for data interpretation, not only nowadays but also in the future.

The advantages and novelties of the developed methodologies have been described in the respective Discussion section of each chapter. The core algorithms of BioInfoMiner have been developed in the past, in the laboratory of Metabolic Engineering and Bioinformatics Program of NHRF, however, within the frames of this thesis, they were redesigned and implemented based on a totally novel software library. Besides some technical improvements in their algorithmic process, one significant advantage is that BioInfoMiner could serve as a integrative pipeline in different computational systems, such as the ANASTASIA platform (chapter 4). Also, defined functions of the software library could be used to analyze and filter the results of BioInfoMiner, customizing the analysis in to the needs of each study, such as in the project described in chapter 6. The background database can be easily enriched with new ontologies and the annotation of new species, as in the study of proteostasis evolutionary imprinting, where more than 400 collections of GO annotations where constructed (chapter 7). The proposed comparative analysis workflow, implemented in chapter 7 and 8 demonstrates the high versatility of the novel software library. That methodology could be used to detect the semantic similarities of the samples of a dataset, targeting to answer in various scientific questions. For example, in chapter 7 it was used to perform a phylogenetic analysis for proteostasis machinery, while in chapter 8 it served as an operator to calculate the functional and phenotypic similarities among the host interactomes of pathogenic viruses. Overall, it is

a novel approach to provide a system-level similarity score for the examined samples, assessing the topologies of their semantic networks.

9.2 Future Work

The suggestions for the improvement of BioInfoMiner and the workflow of comparative analysis could be summarized in the following list:

- Integration of new ontologies with available genomic annotation in BioInfoMiner database (chapter 3).
- Refinement of pathway analysis step with an entropy-based, free of statistical thresholds methodology, able to extract minimum-noise biological information with deeper granularity (chapter 3).
- Development of an additional computational step in the workflow of BioInfoMiner to automatically estimate the regulation of the expression of prioritized genes under chemical and genetic perturbations. L1000 or any other database with gene expression profiles of drug treatments could be used for that task (chapter 8).
- Concerning the ANASTASIA platform (chapter 4), BioTranslator needs to be adapted in the demands of human microbiome studies and another tool to perform metabolic reconstruction from metagenomic data needs to be developed and interconnected with BioTranslator.
- Creation of a new, advanced, free of bias, semantic similarity measure, which would integrate the contribution of the proposed topological factors (chapter 5).
- Integration of the comparative analysis workflow in a backend-frontend environment, such as that of BioInfoMiner, with an additional module for the elucidative and interactive visualization of results.
- Containerization of the developed software library to facilitate its usage in various computational systems without the need to install the requirements of Python and MongoDB.

Appendix A

BioInfoMiner: Interpretation of Cellular Complexity through Semantic Network Analysis

Table A.1: Species-Ontologies Annotations Integrated in BioInfoMiner

Species	Taxonomy	Ontologies
Bos Taurus	Animalia	Gene Ontology Reactome
Caenorhabditis elegans	Animalia	Gene Ontology Reactome WormBase Ontology
Drosophila melanogaster	Animalia	Gene Ontology Reactome
Danio Rerio	Animalia	Gene Ontology Reactome Zebrafish Ontology
Gallus gallus	Animalia	Gene Ontology Reactome

Continued on next page

Table A.1 – continued from previous page

Species	Taxonomy	Ontologies
Homo sapiens	Animalia	Gene Ontology Reactome MGI Mammalian Phenotype Human Phenotype Ontology
Mus musculus	Animalia	Gene Ontology Reactome MGI Mammalian Phenotype
Rattus norvegicus	Animalia	Gene Ontology Reactome MGI Mammalian Phenotype
Sus scrofa	Animalia	Gene Ontology Reactome
Aspergillus nidulans	Fungi	Gene Ontology
Magnaporthe grisea	Fungi	Gene Ontology
Saccharomyces cerevisiae	Fungi	Gene Ontology Reactome
Schizosaccharomyces pombe	Fungi	Gene Ontology Reactome
Arabidopsis thaliana	Plantae	Gene Ontology Plant Ontology
Beta vulgaris	Plantae	Gene Ontology
Glycine max	Plantae	Gene Ontology
Helianthus annuus	Plantae	Gene Ontology
Oryza sativa Japonica	Plantae	Gene Ontology
Sorghum bicolor	Plantae	Gene Ontology
Zea mays	Plantae	Gene Ontology

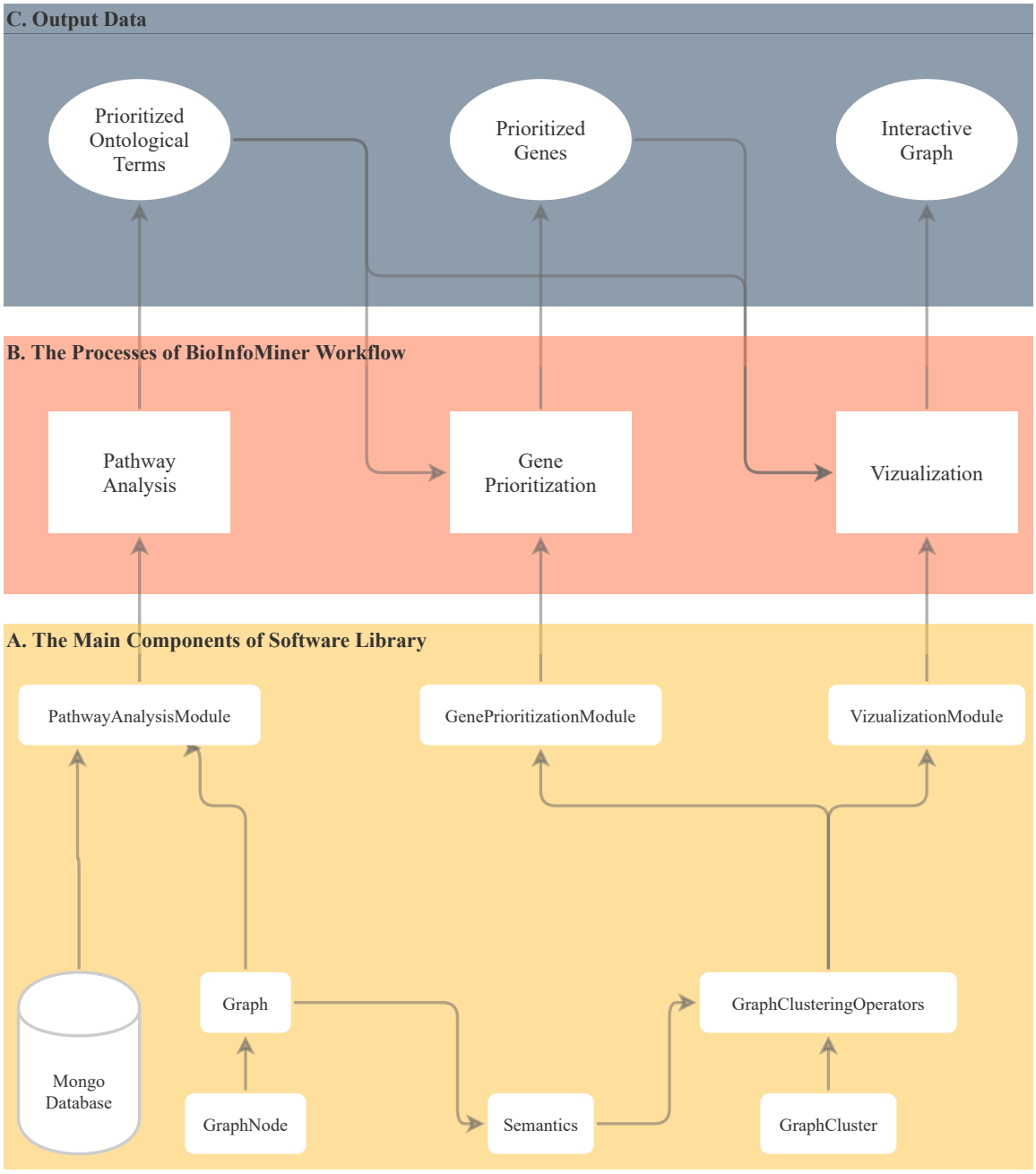


Figure A-1: The detailed workflow of BioInfoMiner. The execution of computational analysis is based on an object-oriented software library, built using Python 2.7. Objects, modules and data stored in MongoDB collections are combined to perform the three processes (pathway analysis, gene prioritization and data visualization) in a sequential mode. Prioritized semantic terms and genes are presented comprehensively in the output interactive graph.

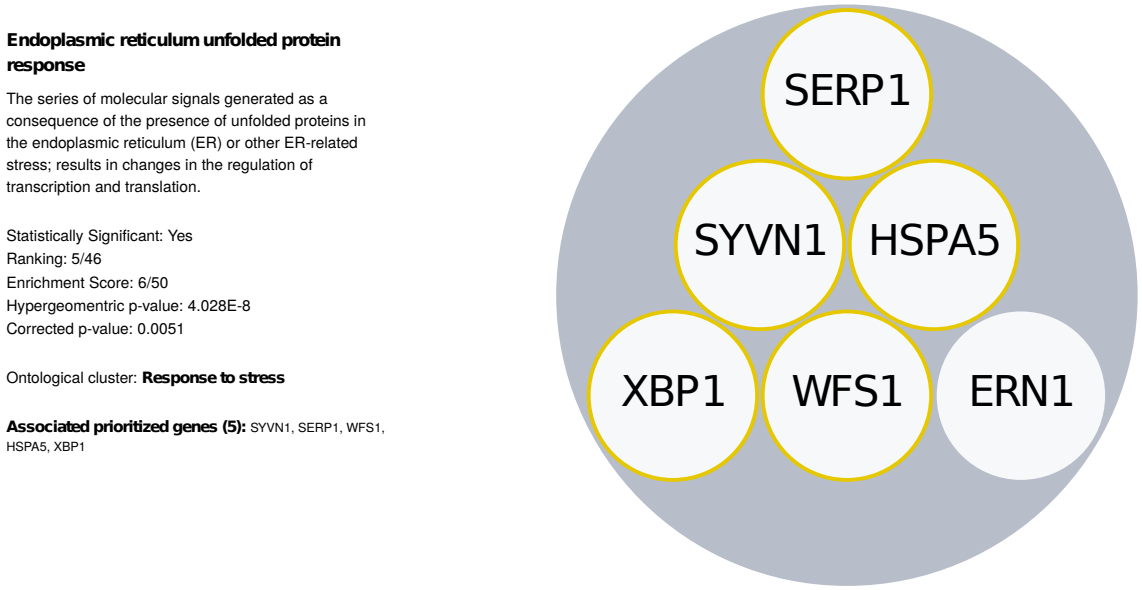


Figure A-2: Graphical representation of the genomic content of an enriched term. The term is presented as a circle packing plot, which includes its associated genes as smaller circles. Genes with gold-colored contour belong to the list of prioritized genes.

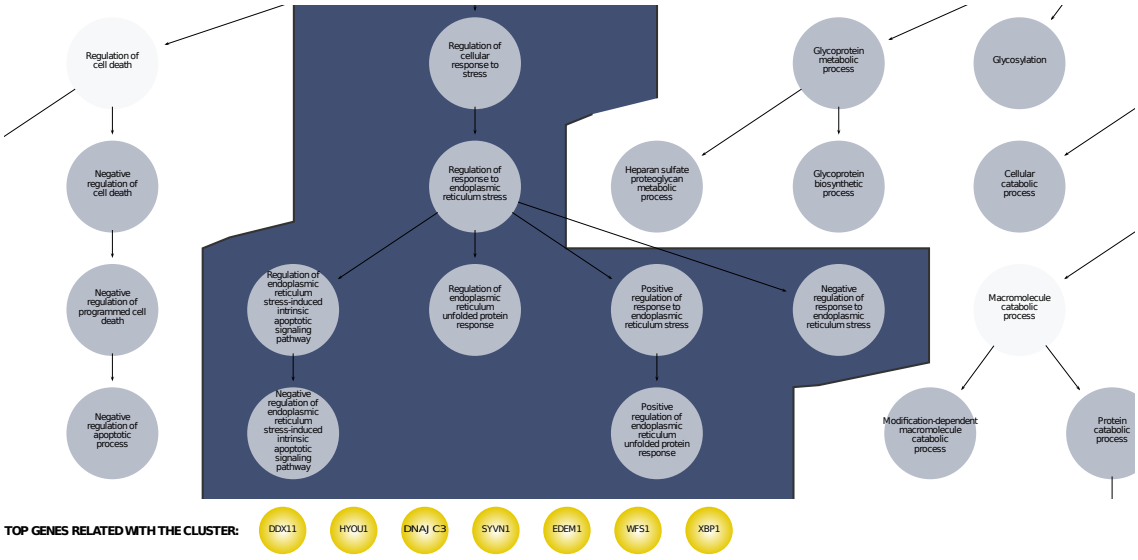


Figure A-3: The user is able to zoom in the content of a specific semantic cluster. The areas around the enriched terms, which belong to the same cluster, shape one or more polygons and their color is uniquely defined. The respective topological boundaries are calculated using the Voronoi algorithm [158]. The bar of gold-colored circles below the main graph illustrates the prioritized genes which have been associated with that cluster.

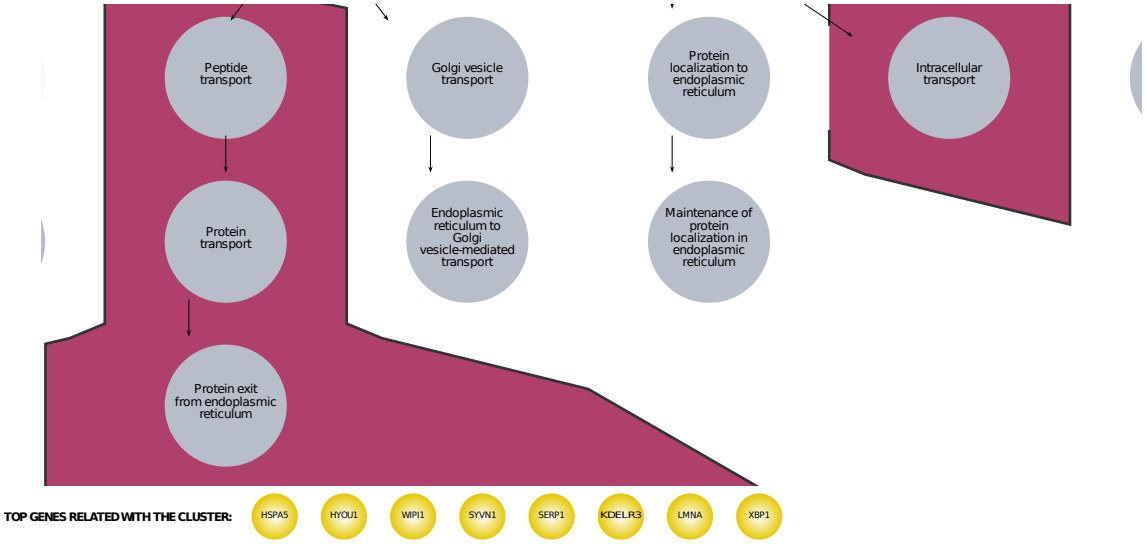


Figure A-4: The user is able to zoom in the content of a specific semantic cluster. The areas around the enriched terms, which belong to the same cluster, shape one or more polygons and their color is uniquely defined. The respective topological boundaries are calculated using the Voronoi algorithm [158]. The bar of gold-colored circles below the main graph illustrates the prioritized genes which have been associated with that cluster.

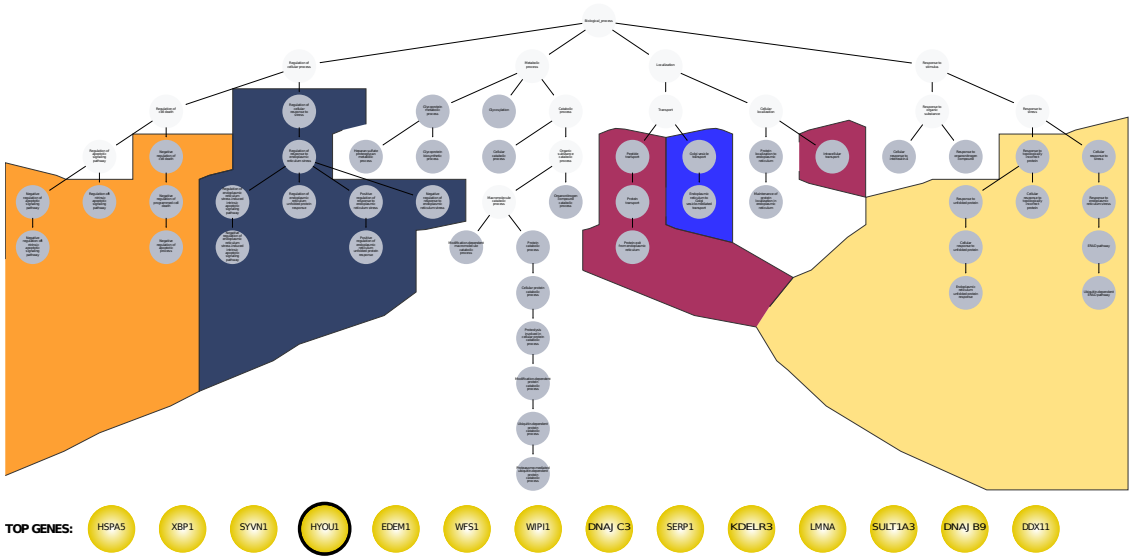


Figure A-5: The user is able to click on a prioritized gene (a gold-colored circle, below the graph) to reveal its semantic profile. The associated clusters are colored uniquely. Their topological boundaries have been defined with the Voronoi algorithm [158].

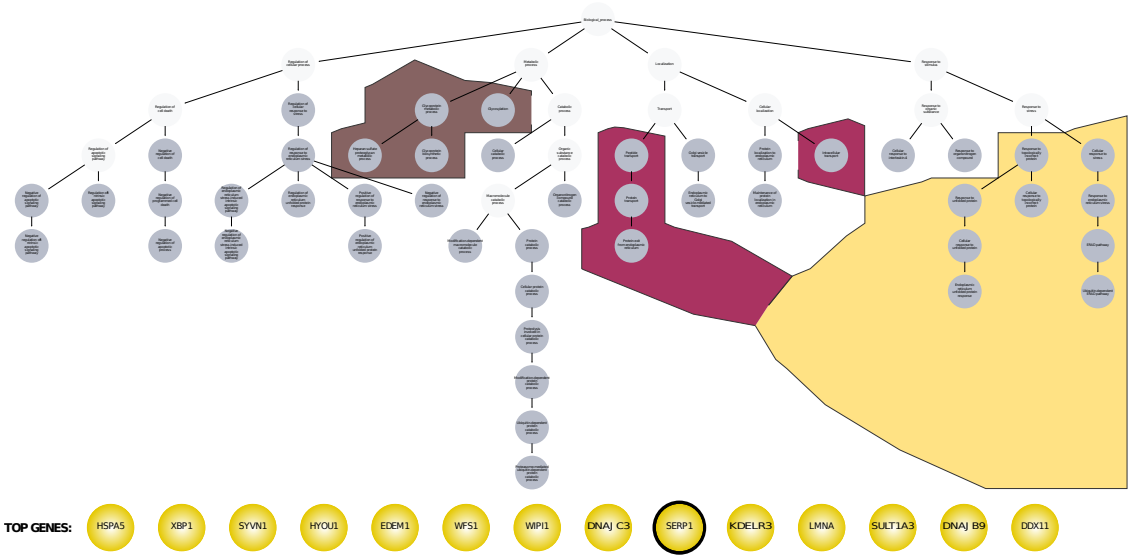


Figure A-6: The user is able to click on a prioritized gene (a gold-colored circle, below the graph) to reveal its semantic profile. The associated clusters are colored uniquely. Their topological boundaries have been defined with the Voronoi algorithm [158].

Appendix B

Case Study 1: Profiling of TRAIL-Induced Modulation of NK cells During Viral Infection

Table B.1: Enriched GO-BP Terms for the Wild Type (WT) Activation

Rank	Biological Process	Enrichment	Adjusted p-value
1	Response to virus	65/248	0.0005
2	Response to interferon-beta	27/45	0.0013
3	Defense response to virus	52/177	0.0018
4	Response to other organism	136/977	0.0019
5	Response to external biotic stimulus	136/980	0.0022
6	Cellular response to interferon-beta	23/37	0.0029
7	Response to biotic stimulus	139/1020	0.0032
8	Defense response to other organism	93/564	0.0042
9	Immune response	143/1179	0.0050
10	Cellular response to cytokine stimulus	96/704	0.0050
11	Immune effector process	82/532	0.0052
12	Defense response	158/1241	0.0055
13	Response to cytokine	110/821	0.0061
14	Response to bacterium	90/740	0.0073
15	Immune system process	201/1969	0.0074
16	Innate immune response	92/671	0.0075
17	Response to external stimulus	198/2000	0.0078
18	Response to interferon-alpha	13/25	0.0085
19	Regulation of cytokine production	74/586	0.0087

Continued on next page

Table B.1 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
20	Cellular response to organic substance	175/1914	0.0087
21	Cellular response to chemical stimulus	209/2346	0.0095
22	Regulation of defense response	63/537	0.0115
23	Defense response to protozoan	12/29	0.0118
24	Regulation of immune system process	127/1289	0.0119
25	Regulation of cytokine production involved in immune response	17/70	0.0133
26	Negative regulation of viral genome replication	14/42	0.0134
27	Response to protozoan	12/32	0.0140
28	Response to stress	252/3089	0.0143
29	Positive regulation of defense response	40/270	0.0145
30	Regulation of immune response	80/748	0.0148
31	Response to interferon-gamma	24/129	0.0149
32	Negative regulation of viral life cycle	16/65	0.0157
33	Negative regulation of viral process	18/81	0.0163
34	Positive regulation of innate immune response	27/160	0.0165
35	Negative regulation of cytokine production	33/220	0.0174
36	Regulation of multi-organism process	47/368	0.0175
37	Response to type I interferon	8/16	0.0184
38	Positive regulation of immune system process	88/877	0.0207
39	Negative regulation of cytokine production involved in immune response	9/23	0.0210
40	Negative regulation of multi-organism process	27/164	0.0211
41	Negative regulation of immune response	23/140	0.0214
42	Inflammatory response	51/435	0.0217
43	Cytokine-mediated signaling pathway	42/333	0.0219
44	Adaptive immune response	45/375	0.0223
45	Regulation of immune effector process	43/340	0.0224
46	Alpha-beta T cell activation	15/65	0.0228
47	Cellular response to lipopolysaccharide	25/160	0.0228
48	Regulation of innate immune response	33/241	0.0230
49	Defense response to Gram-positive bacterium	17/89	0.0260
50	Cellular response to interferon-gamma	19/105	0.0267
51	T cell activation	33/244	0.0268
52	Regulation of viral process	24/157	0.0270
53	Alpha-beta T cell differentiation	13/55	0.0279
54	Response to lipopolysaccharide	39/316	0.0279
55	Cell surface receptor signaling pathway	152/1802	0.0279
56	Cellular response to molecule of bacterial origin	25/166	0.0287

Continued on next page

Table B.1 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
57	Cellular response to biotic stimulus	27/186	0.0289
58	Response to molecule of bacterial origin	40/332	0.0291
59	Positive regulation of cell death	66/663	0.0294
60	Regulation of viral genome replication	15/75	0.0297
61	Multi-organism process	174/2115	0.0299
62	Positive regulation of cytokine production	44/379	0.0300
63	Positive regulation of immune response	57/555	0.0306
64	Negative regulation of immune system process	48/441	0.0307
65	Defense response to bacterium	43/371	0.0311
66	Cellular response to interferon-alpha	6/13	0.0322
67	Positive regulation of response to stimulus	164/2031	0.0332
68	Regulation of symbiosis, encompassing mutualism through parasitism	27/192	0.0335
69	Regulation of leukocyte cell-cell adhesion	32/262	0.0349
70	Regulation of viral life cycle	19/122	0.0352
71	Regulation of response to stress	104/1177	0.0355
72	Regulation of cell adhesion	61/615	0.0356
73	Negative regulation of production of molecular mediator of immune response	9/31	0.0358
74	Negative regulation of cytokine biosynthetic process	8/26	0.0364
75	Negative regulation of lymphocyte mediated immunity	10/41	0.0364
76	Positive regulation of programmed cell death	60/610	0.0381
77	Regulation of T cell cytokine production	8/28	0.0385
78	Regulation of lymphocyte mediated immunity	21/142	0.0387
79	Regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	21/146	0.0396
80	Positive regulation of smooth muscle cell proliferation	17/107	0.0397
81	Negative regulation of cell activation	24/178	0.0405
82	Positive regulation of apoptotic process	59/605	0.0414
83	Negative regulation of leukocyte mediated immunity	11/51	0.0418
84	Cyclooxygenase pathway	3/3	0.0418
85	Negative regulation of lymphocyte activation	20/137	0.0419
86	Regulation of smooth muscle cell proliferation	22/165	0.0423
87	Regulation of adaptive immune response	22/158	0.0425
88	Regulation of T cell mediated immunity	12/61	0.0432
89	Negative regulation of T cell activation	17/108	0.0433

Continued on next page

Table B.1 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
90	Negative regulation of leukocyte activation	22/160	0.0433
91	Antigen processing and presentation	14/81	0.0435
92	Regulation of cell activation	59/620	0.0441
93	Regulation of production of molecular mediator of immune response	19/130	0.0444
94	Angiogenesis	33/284	0.0452
95	Negative regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	9/37	0.0456
96	T cell differentiation	20/145	0.0459
97	Regulation of response to external stimulus	63/675	0.0473
98	Positive regulation of T cell proliferation	15/93	0.0475
99	Apoptotic signaling pathway	31/270	0.0497
100	Regulation of vascular smooth muscle cell proliferation	11/56	0.0500

Table B.2: Enriched Reactome Terms for the Wild Type (WT) Activation

Rank	Reactome Pathway	Enrichment	Adjusted p-value
1	Immune System	145/1809	0.0016
2	Interferon Signaling	17/67	0.0039
3	Cytokine Signaling in Immune system	46/430	0.0067
4	Interferon alpha/beta signaling	10/30	0.0105
5	Regulation of IFNA signaling	8/25	0.0115
6	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	18/113	0.0125
7	Nicotinamide salvaging	7/20	0.0136
8	Extracellular matrix organization	32/323	0.0190
9	Nicotinate metabolism	8/32	0.0228
10	Integrin cell surface interactions	13/80	0.0228
11	Downstream signal transduction	7/29	0.0269
12	Neurophilin interactions with VEGF and VEGFR	3/4	0.0279
13	Signaling by PDGF	10/58	0.0285
14	Adaptive Immune System	58/766	0.0319
15	Hemostasis	48/616	0.0366
16	Platelet activation, signaling and aggregation	25/266	0.0383
17	Repression of WNT target genes	4/9	0.0397
18	Role of phospholipids in phagocytosis	10/66	0.0441
19	Non-integrin membrane-ECM interactions	7/38	0.0450

Continued on next page

Table B.2 – continued from previous page

Rank	Reactome Pathway	Enrichment	Adjusted p-value
20	Antigen processing-Cross presentation	11/90	0.0454
21	Metabolism of water-soluble vitamins and cofactors	14/123	0.0473
22	FCGR activation	8/57	0.0501
23	Platelet degranulation	14/138	0.0540
24	Signaling by Interleukins	25/305	0.0559
25	Antigen Presentation: Folding, assembly and peptide loading of class I MHC	6/39	0.0606
26	Nitric oxide stimulates guanylate cyclase	4/17	0.0607
27	Response to elevated platelet cytosolic Ca2+	14/143	0.0623
28	Neutrophil degranulation	40/567	0.0623
29	Interleukin-9 signaling	3/10	0.0635
30	Metabolism of vitamins and cofactors	17/193	0.0710
31	Other interleukin signaling	4/19	0.0713
32	Chemokine receptors bind chemokines	7/52	0.0722
33	Termination of O-glycan biosynthesis	4/20	0.0758
34	Interleukin-2 family signaling	6/43	0.0778
35	Innate Immune System	66/1061	0.0800
36	Platelet homeostasis	8/72	0.0819
37	TNFR2 non-canonical NF-kB pathway	10/100	0.0833
38	CD28 dependent PI3K/Akt signaling	4/22	0.0844
39	Elevation of cytosolic Ca2+ levels	3/13	0.0866
40	Branched-chain amino acid catabolism	4/23	0.0897
41	Binding and Uptake of Ligands by Scavenger Receptors	8/79	0.0912
42	Metabolism	107/1852	0.0935

Table B.3: Enriched GO-BP Terms for the TRAIL Knock Out (KO) Activation

Rank	Biological Process	Enrichment	Adjusted p-value
1	Response to virus	67/248	0.0002
2	Defense response to virus	52/177	0.0008
3	Response to cytokine	128/821	0.0017
4	Response to interferon-beta	25/45	0.0022
5	Defense response	163/1241	0.0022
6	Cellular response to cytokine stimulus	111/704	0.0023
7	Response to other organism	133/977	0.0031
8	Cellular response to interferon-beta	21/37	0.0037
9	Response to external biotic stimulus	133/980	0.0039
10	Immune response	151/1179	0.0047

Continued on next page

Table B.3 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
11	Immune system process	217/1969	0.0058
12	Innate immune response	92/671	0.0061
13	Response to biotic stimulus	135/1020	0.0063
14	Response to external stimulus	211/2000	0.0072
15	Regulation of defense response	78/537	0.0074
16	Immune effector process	77/532	0.0077
17	Defense response to other organism	80/564	0.0078
18	Regulation of cytokine production	83/586	0.0081
19	Response to interferon-alpha	13/25	0.0096
20	Response to interferon-gamma	30/129	0.0096
21	Cellular response to chemical stimulus	234/2346	0.0104
22	Positive regulation of defense response	48/270	0.0106
23	Inflammatory response	63/435	0.0110
24	Regulation of response to external stimulus	88/675	0.0114
25	Negative regulation of viral genome replication	15/42	0.0120
26	Cytokine-mediated signaling pathway	52/333	0.0122
27	Regulation of response to stress	129/1177	0.0128
28	Cellular response to organic substance	197/1914	0.0130
29	Defense response to protozoan	12/29	0.0141
30	Regulation of immune effector process	51/340	0.0145
31	Regulation of multi-organism process	53/368	0.0147
32	Response to bacterium	88/740	0.0156
33	Regulation of immune system process	135/1289	0.0159
34	Negative regulation of viral process	20/81	0.0160
35	Response to type I interferon	9/16	0.0161
36	Positive regulation of innate immune response	30/160	0.0164
37	Response to protozoan	12/32	0.0165
38	Negative regulation of cytokine production	37/220	0.0168
39	Negative regulation of multi-organism process	30/164	0.0181
40	Response to lipopolysaccharide	46/316	0.0188
41	Negative regulation of viral life cycle	17/65	0.0200
42	Positive regulation of response to external stimulus	42/276	0.0203
43	Regulation of innate immune response	38/241	0.0217
44	Cellular response to interferon-gamma	22/105	0.0219
45	Regulation of cytokine production involved in immune response	18/70	0.0220
46	Positive regulation of response to cytokine stimulus	14/48	0.0226
47	Positive regulation of cytokine production	51/379	0.0229
48	Chemotaxis	54/424	0.0230

Continued on next page

Table B.3 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
49	Regulation of defense response to virus by host	12/39	0.0232
50	Response to molecule of bacterial origin	47/332	0.0234
51	Positive regulation of immune system process	96/877	0.0236
52	Cellular response to lipopolysaccharide	28/160	0.0247
53	Leukocyte chemotaxis	23/116	0.0248
54	Positive regulation of cell death	75/663	0.0262
55	Regulation of interferon-alpha production	10/26	0.0270
56	Taxis	54/426	0.0276
57	Regulation of production of molecular mediator of immune response	23/130	0.0276
58	Cellular response to molecule of bacterial origin	28/166	0.0278
59	Leukocyte migration	30/186	0.0280
60	Positive regulation of cytokine-mediated signaling pathway	12/41	0.0280
61	Regulation of symbiosis, encompassing mutualism through parasitism	30/192	0.0285
62	Negative regulation of immune response	24/140	0.0288
63	Positive regulation of programmed cell death	69/610	0.0295
64	Regulation of immune response	81/748	0.0296
65	Regulation of response to biotic stimulus	24/135	0.0302
66	Cellular response to biotic stimulus	29/186	0.0306
67	Alpha-beta T cell activation	15/65	0.0308
68	Negative regulation of lymphocyte mediated immunity	11/41	0.0311
69	Regulation of interferon-gamma production	18/95	0.0321
70	Positive regulation of apoptotic process	68/605	0.0326
71	Regulation of viral genome replication	16/75	0.0328
72	T cell migration	8/20	0.0339
73	Positive regulation of defense response to virus by host	9/27	0.0350
74	Cell chemotaxis	27/177	0.0350
75	Positive regulation of interferon-alpha production	8/21	0.0353
76	Response to organic substance	240/2817	0.0387
77	Positive regulation of inflammatory response	20/113	0.0391
78	Alpha-beta T cell differentiation	13/55	0.0392
79	Regulation of cell adhesion	67/615	0.0393
80	Lymphocyte migration	13/56	0.0395
81	Regulation of viral process	25/157	0.0400
82	Response to stress	261/3089	0.0400

Continued on next page

Table B.3 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted p-value
83	Negative regulation of leukocyte mediated immunity	12/51	0.0403
84	Positive regulation of smooth muscle cell proliferation	19/107	0.0405
85	Multi-organism process	187/2115	0.0405
86	Regulation of T cell activation	37/286	0.0405
87	Cellular response to virus	9/30	0.0410
88	Positive regulation of intracellular signal transduction	92/920	0.0414
89	Negative regulation of innate immune response	12/50	0.0417
90	Cellular response to interferon-alpha	6/13	0.0421
91	Regulation of viral life cycle	20/122	0.0433
92	Regulation of smooth muscle cell proliferation	25/165	0.0444
93	Positive regulation of cell adhesion	44/368	0.0444
94	Myeloid leukocyte migration	18/101	0.0445
95	Positive regulation of response to stimulus	179/2031	0.0448
96	Regulation of leukocyte proliferation	30/220	0.0460
97	Regulation of response to cytokine stimulus	19/110	0.0462
98	Regulation of inflammatory response	37/292	0.0464
99	Regulation of leukocyte cell-cell adhesion	34/262	0.0465
100	T cell activation	32/244	0.0477
101	Cellular response to amyloid-beta	8/25	0.0491
102	Cell death	85/859	0.0492
103	Negative regulation of T cell activation	18/108	0.0500

Table B.4: Enriched Reactome Terms for the TRAIL Knock Out (KO) Activation

Rank	Reactome Pathway	Enrichment	Adjusted p-value
1	Immune System	164/1809	0.0025
2	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	24/113	0.0049
3	Cytokine Signaling in Immune system	53/430	0.0061
4	Interferon Signaling	18/67	0.0085
5	Interferon alpha/beta signaling	9/30	0.0096
6	Regulation of IFNA signaling	8/25	0.0119
7	Interleukin-6 signaling	5/11	0.0157
8	Adaptive Immune System	62/766	0.0178
9	Extracellular matrix organization	32/323	0.0216
10	Neutrophil degranulation	48/567	0.0221
11	Nicotinamide salvaging	6/20	0.0241

Continued on next page

Table B.4 – continued from previous page

Rank	Reactome Pathway	Enrichment	Adjusted p-value
12	Downstream signal transduction	7/29	0.0280
13	Eicosanoid ligand-binding receptors	5/14	0.0286
14	Signaling by PDGF	10/58	0.0296
15	Innate Immune System	78/1061	0.0327
16	Chemokine receptors bind chemokines	9/52	0.0369
17	Hemostasis	49/616	0.0387
18	Interleukin-12 family signaling	5/20	0.0388
19	Signaling by Interleukins	28/305	0.0391
20	Integrin cell surface interactions	11/80	0.0438
21	Metabolism of amino acids and derivatives	23/254	0.0453
22	Class A/1 (Rhodopsin-like receptors)	28/318	0.0455
23	TNFR2 non-canonical NF-kB pathway	12/100	0.0474
24	Platelet activation, signaling and aggregation	24/266	0.0512
25	Metal sequestration by antimicrobial proteins	2/3	0.0546
26	Transport of organic anions	4/15	0.0564
27	Generation of second messenger molecules	5/26	0.0625
28	GPCR ligand binding	33/424	0.0638
29	SLC-mediated transmembrane transport	22/250	0.0641
30	Collagen degradation	9/71	0.0644
31	Post-translational protein phosphorylation	13/123	0.0657
32	Transport of small molecules	50/705	0.0697
33	PD-1 signaling	4/19	0.0714
34	Non-integrin membrane-ECM interactions	6/38	0.0773
35	Signaling by SCF-KIT	6/40	0.0781
36	Negative regulation of the PI3K/AKT network	11/104	0.0792
37	Transport of vitamins, nucleosides, and related molecules	6/41	0.0809
38	Termination of O-glycan biosynthesis	4/20	0.0826
39	Programmed Cell Death	11/105	0.0860
40	Transport of inorganic cations/anions and amino acids/oligopeptides	11/109	0.0883
41	Metabolism	113/1852	0.0893
42	Reversible hydration of carbon dioxide	3/12	0.0916
43	Signaling by NOTCH	6/45	0.0945
44	Platelet degranulation	13/138	0.0954
45	Amino acid transport across the plasma membrane	5/33	0.0978
46	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	10/97	0.0982

Appendix C

Case Study 2: Proteostasis Imprinting Across Evolution

Table C.1: Species Used to Reveal the Proteostasis Profile Across Taxonomies

Species	Superkingdom	Phylum	Class
Tribolium castaneum	Eukaryotes	Arthropoda	Insecta
Solenopsis invicta	Eukaryotes	Arthropoda	Insecta
Rhodnius prolixus	Eukaryotes	Arthropoda	Insecta
Pediculus humanus subsp. corporis	Eukaryotes	Arthropoda	Insecta
Drosophila simulans	Eukaryotes	Arthropoda	Insecta
Drosophila melanogaster	Eukaryotes	Arthropoda	Insecta
Bombyx mori	Eukaryotes	Arthropoda	Insecta
Anopheles gambiae	Eukaryotes	Arthropoda	Insecta
Anopheles darlingi	Eukaryotes	Arthropoda	Insecta
Saccharomyces cerevisiae	Eukaryotes	Ascomycota	Saccharomycetes
Komagataella pastoris	Eukaryotes	Ascomycota	Saccharomycetes
Schizosaccharomyces pombe	Eukaryotes	Ascomycota	Schizosaccharomycetes
Schizosaccharomyces octosporus	Eukaryotes	Ascomycota	Schizosaccharomycetes
Schizosaccharomyces japonicus	Eukaryotes	Ascomycota	Schizosaccharomycetes
Puccinia graminis	Eukaryotes	Basidiomycota	Pucciniomycetes
Cryptococcus neoformans	Eukaryotes	Basidiomycota	Tremellomycetes
Sporisorium reilianum	Eukaryotes	Basidiomycota	Ustilaginomycetes
Chlamydomonas reinhardtii	Eukaryotes	Chlorophyta	Chlorophyceae
Xiphophorus maculatus	Eukaryotes	Chordata	Actinopteri
Scophthalmus maximus	Eukaryotes	Chordata	Actinopteri
Poecilia reticulata	Eukaryotes	Chordata	Actinopteri
Oryzias latipes	Eukaryotes	Chordata	Actinopteri
Lates calcarifer	Eukaryotes	Chordata	Actinopteri
Labrus bergylta	Eukaryotes	Chordata	Actinopteri
Gadus morhua	Eukaryotes	Chordata	Actinopteri
Fundulus heteroclitus	Eukaryotes	Chordata	Actinopteri
Danio rerio	Eukaryotes	Chordata	Actinopteri
Cyprinodon variegatus	Eukaryotes	Chordata	Actinopteri
Clupea harengus	Eukaryotes	Chordata	Actinopteri
Xenopus tropicalis	Eukaryotes	Chordata	Amphibia
Ciona intestinalis	Eukaryotes	Chordata	Ascidiacea
Melopsittacus undulatus	Eukaryotes	Chordata	Aves
Meleagris gallopavo	Eukaryotes	Chordata	Aves
Gallus gallus	Eukaryotes	Chordata	Aves
Dromaius novaehollandiae	Eukaryotes	Chordata	Aves

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Anas platyrhynchos	Eukaryotes	Chordata	Aves
Callorhinchus milii	Eukaryotes	Chordata	Chondrichthyes
Erpetoichthys calabaricus	Eukaryotes	Chordata	Cladistia
Latimeria chalumnae	Eukaryotes	Chordata	Coelacanthi
Vombatus ursinus	Eukaryotes	Chordata	Mammalia
Sus scrofa	Eukaryotes	Chordata	Mammalia
Rattus norvegicus	Eukaryotes	Chordata	Mammalia
Pongo abelii	Eukaryotes	Chordata	Mammalia
Pan troglodytes	Eukaryotes	Chordata	Mammalia
Ovis aries	Eukaryotes	Chordata	Mammalia
Oryctolagus cuniculus	Eukaryotes	Chordata	Mammalia
Ornithorhynchus anatinus	Eukaryotes	Chordata	Mammalia
Nomascus leucogenys	Eukaryotes	Chordata	Mammalia
Microtus ochrogaster	Eukaryotes	Chordata	Mammalia
Macaca mulatta	Eukaryotes	Chordata	Mammalia
Homo sapiens	Eukaryotes	Chordata	Mammalia
Gorilla gorilla	Eukaryotes	Chordata	Mammalia
Felis catus	Eukaryotes	Chordata	Mammalia
Erinaceus europaeus	Eukaryotes	Chordata	Mammalia
Equus caballus	Eukaryotes	Chordata	Mammalia
Dasypus novemcinctus	Eukaryotes	Chordata	Mammalia
Callithrix jacchus	Eukaryotes	Chordata	Mammalia
Bos taurus	Eukaryotes	Chordata	Mammalia
Ailuropoda melanoleuca	Eukaryotes	Chordata	Mammalia
Sphenodon punctatus	Eukaryotes	Chordata	Reptilia
Pelodiscus sinensis	Eukaryotes	Chordata	Reptilia
Crocodylus porosus	Eukaryotes	Chordata	Reptilia
Anolis carolinensis	Eukaryotes	Chordata	Reptilia
Crassostrea gigas	Eukaryotes	Mollusca	Bivalvia
Lepeophtheirus salmonis	Eukaryotes	Negarnaviricota	Monjiviricetes
Strongyloides ratti	Eukaryotes	Nematoda	Chromadorea
Loa loa	Eukaryotes	Nematoda	Chromadorea
Caenorhabditis elegans	Eukaryotes	Nematoda	Chromadorea
Caenorhabditis briggsae	Eukaryotes	Nematoda	Chromadorea
Trichinella spiralis	Eukaryotes	Nematoda	Enoplea
Schistosoma mansoni	Eukaryotes	Platyhelminthes	Trematoda
Galdieria sulphuraria	Eukaryotes	Rhodophyta	Bangiophyceae
Cyanidioschyzon merolae	Eukaryotes	Rhodophyta	Bangiophyceae
Zea mays	Eukaryotes	Streptophyta	Liliopsida
Oryza sativa Japonica Group	Eukaryotes	Streptophyta	Liliopsida
Oryza punctata	Eukaryotes	Streptophyta	Liliopsida
Oryza glumipatula	Eukaryotes	Streptophyta	Liliopsida
Oryza glaberrima	Eukaryotes	Streptophyta	Liliopsida
Oryza barthii	Eukaryotes	Streptophyta	Liliopsida
Musa acuminata	Eukaryotes	Streptophyta	Liliopsida
Vitis vinifera	Eukaryotes	Streptophyta	Magnoliopsida
Solanum tuberosum	Eukaryotes	Streptophyta	Magnoliopsida
Solanum lycopersicum	Eukaryotes	Streptophyta	Magnoliopsida
Nicotiana attenuata	Eukaryotes	Streptophyta	Magnoliopsida
Medicago truncatula	Eukaryotes	Streptophyta	Magnoliopsida
Helianthus annuus	Eukaryotes	Streptophyta	Magnoliopsida
Glycine max	Eukaryotes	Streptophyta	Magnoliopsida
Cucumis sativus	Eukaryotes	Streptophyta	Magnoliopsida
Capsicum annuum	Eukaryotes	Streptophyta	Magnoliopsida
Brassica rapa	Eukaryotes	Streptophyta	Magnoliopsida
Brassica oleracea	Eukaryotes	Streptophyta	Magnoliopsida
Brassica napus	Eukaryotes	Streptophyta	Magnoliopsida
Arabidopsis thaliana	Eukaryotes	Streptophyta	Magnoliopsida
Amborella trichopoda	Eukaryotes	Streptophyta	Magnoliopsida
Abditibacterium utsteinense	Bacteria	Abditibacteriota	Abditibacteria
Terriglobus saanensis	Bacteria	Acidobacteria	Acidobacteriia
Terriglobus roseus	Bacteria	Acidobacteria	Acidobacteriia

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Granulicella rosea	Bacteria	Acidobacteria	Acidobacteriia
Luteitalea pratensis	Bacteria	Acidobacteria	Vicinamibacteria
Streptomyces sp NRRL S-495	Bacteria	Actinobacteria	Actinobacteria
Streptomyces sp CC53	Bacteria	Actinobacteria	Actinobacteria
Streptomyces shenzhenensis	Bacteria	Actinobacteria	Actinobacteria
Streptomyces jeddahensis	Bacteria	Actinobacteria	Actinobacteria
Streptomyces avermitilis	Bacteria	Actinobacteria	Actinobacteria
Rhodococcus sp	Bacteria	Actinobacteria	Actinobacteria
Rhodococcus coprophilus	Bacteria	Actinobacteria	Actinobacteria
Mycolicibacterium holsaticum	Bacteria	Actinobacteria	Actinobacteria
Mycobacterium tuberculosis	Bacteria	Actinobacteria	Actinobacteria
Mycobacterium marinum	Bacteria	Actinobacteria	Actinobacteria
Micromonospora nigra	Bacteria	Actinobacteria	Actinobacteria
Microbacterium hydrocarbonoxydans	Bacteria	Actinobacteria	Actinobacteria
Leucobacter massiliensis	Bacteria	Actinobacteria	Actinobacteria
Leucobacter chromiiresistens	Bacteria	Actinobacteria	Actinobacteria
Blastococcus sp TF02A-30	Bacteria	Actinobacteria	Actinobacteria
Blastococcus sp CT_GayMR16	Bacteria	Actinobacteria	Actinobacteria
Blastococcus saxobsidens	Bacteria	Actinobacteria	Actinobacteria
Bifidobacterium cuniculi	Bacteria	Actinobacteria	Actinobacteria
Bifidobacterium bohemicum	Bacteria	Actinobacteria	Actinobacteria
Amycolatopsis keratiniphila	Bacteria	Actinobacteria	Actinobacteria
Actinosynnema mirum	Bacteria	Actinobacteria	Actinobacteria
Actinoplanes sp	Bacteria	Actinobacteria	Actinobacteria
Actinomyces tangfeifanii	Bacteria	Actinobacteria	Actinobacteria
Actinomyces howellii	Bacteria	Actinobacteria	Actinobacteria
Actinoalloteichus hoggarensis	Bacteria	Actinobacteria	Actinobacteria
Parvibacter caecicola	Bacteria	Actinobacteria	Coriobacteriia
Egibacter rhizosphaerae	Bacteria	Actinobacteria	Nitriliruptoria
Rubrobacter xylanophilus	Bacteria	Actinobacteria	Rubrobacteria
Conexibacter woesei	Bacteria	Actinobacteria	Thermoleophilia
Aquifex aeolicus	Bacteria	Aquificae	Aquificae
Salinivirga cyanobacteriivorans	Bacteria	Bacteroidetes	Bacteroidia
Proteiniphilum saccharofermentans	Bacteria	Bacteroidetes	Bacteroidia
Prevotella sp	Bacteria	Bacteroidetes	Bacteroidia
Prevotella baroniae	Bacteria	Bacteroidetes	Bacteroidia
Paramuribaculum intestinale	Bacteria	Bacteroidetes	Bacteroidia
Dysgonomonas capnocytophagoides	Bacteria	Bacteroidetes	Bacteroidia
Bacteroides stercoris	Bacteria	Bacteroidetes	Bacteroidia
Bacteroides faecichinchillae	Bacteria	Bacteroidetes	Bacteroidia
Bacteroides caccae	Bacteria	Bacteroidetes	Bacteroidia
Flaviaesturariibacter sp	Bacteria	Bacteroidetes	Chitinophagia
Chitinophaga pinensis	Bacteria	Bacteroidetes	Chitinophagia
Chitinophaga niabensis	Bacteria	Bacteroidetes	Chitinophagia
Spirosoma radiotolerans	Bacteria	Bacteroidetes	Cytophagia
Pontibacter sp	Bacteria	Bacteroidetes	Cytophagia
Marivirga tractuosa	Bacteria	Bacteroidetes	Cytophagia
Hymenobacter sedentarius	Bacteria	Bacteroidetes	Cytophagia
Algoriphagus machipongonensis	Bacteria	Bacteroidetes	Cytophagia
Robertkochia marina	Bacteria	Bacteroidetes	Flavobacteriia
Psychroflexus halocasei	Bacteria	Bacteroidetes	Flavobacteriia
Flavobacterium succinicans	Bacteria	Bacteroidetes	Flavobacteriia
Flavobacterium sp	Bacteria	Bacteroidetes	Flavobacteriia
Cruoricaptor ignavus	Bacteria	Bacteroidetes	Flavobacteriia
Croceitalea dokdonensis	Bacteria	Bacteroidetes	Flavobacteriia
Chryseobacterium nematophagum	Bacteria	Bacteroidetes	Flavobacteriia
Arenibacter sp	Bacteria	Bacteroidetes	Flavobacteriia
Saprospira grandis	Bacteria	Bacteroidetes	Saprospiria
Sphingobacterium sp	Bacteria	Bacteroidetes	Sphingobacteriia
Pedobacter sp	Bacteria	Bacteroidetes	Sphingobacteriia
Pedobacter psychrophilus	Bacteria	Bacteroidetes	Sphingobacteriia
Pedobacter luteus	Bacteria	Bacteroidetes	Sphingobacteriia

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Mucilaginibacter sp ZH6	Bacteria	Bacteroidetes	Sphingobacteriia
Mucilaginibacter sp PPCGB 2223	Bacteria	Bacteroidetes	Sphingobacteriia
Mucilaginibacter lappiensis	Bacteria	Bacteroidetes	Sphingobacteriia
Rhodohalobacter barkolensis	Bacteria	Balneolaeota	Balneolia
Gracilimonas sp	Bacteria	Balneolaeota	Balneolia
Caldithrix abyssi	Bacteria	Calditrichaeota	Calditrichae
Parachlamydia acanthamoebae	Bacteria	Chlamydiae	Chlamydiia
Prosthecochloris sp	Bacteria	Chlorobi	Chlorobia
Pelolinea submarina	Bacteria	Chloroflexi	Anaerolineae
Ardenticatena maritima	Bacteria	Chloroflexi	Ardenticatenia
Caldilinea aerophila	Bacteria	Chloroflexi	Caldilineae
Roseiflexus sp	Bacteria	Chloroflexi	Chloroflexia
Dehalogenimonas lykanthroporepellens	Bacteria	Chloroflexi	Dehalococcoidia
Ktedonobacterales bacterium	Bacteria	Chloroflexi	Ktedonobacteria
Thermoflexus hugenholtzii	Bacteria	Chloroflexi	Thermoflexia
Nitrolancea hollandica Lb	Bacteria	Chloroflexi	Thermomicrobia
Desulfurispirillum indicum	Bacteria	Chrysiogenetes	Chrysiogenetes
Nostoc sp	Bacteria	Cyanobacteria	Cyanophyceae
Leptolyngbya sp	Bacteria	Cyanobacteria	Cyanophyceae
Gloeobacter violaceus	Bacteria	Cyanobacteria	Gloeobacteria
Geovibrio thiophilus	Bacteria	Deferribacteres	Deferribacteres
Oceanithermus profundus	Bacteria	Deinococcus-Thermus	Deinococci
Meiothermus silvanus	Bacteria	Deinococcus-Thermus	Deinococci
Marinithermus hydrothermalis	Bacteria	Deinococcus-Thermus	Deinococci
Deinococcus sp	Bacteria	Deinococcus-Thermus	Deinococci
Dictyoglomus turgidum	Bacteria	Dictyoglomi	Dictyoglomia
Elusimicrobium sp An273	Bacteria	Elusimicrobia	Elusimicrobia
Endomicrobium proavitum	Bacteria	Elusimicrobia	Endomicrobia
Fibrobacter sp UWB8	Bacteria	Fibrobacteres	Fibrobacteria
Streptococcus pneumoniae	Bacteria	Firmicutes	Bacilli
Staphylococcus epidermidis	Bacteria	Firmicutes	Bacilli
Listeria monocytogenes serovar 1/2a	Bacteria	Firmicutes	Bacilli
Lactococcus lactis subsp lactis	Bacteria	Firmicutes	Bacilli
Brevibacillus brevis	Bacteria	Firmicutes	Bacilli
Bacillus taeanensis	Bacteria	Firmicutes	Bacilli
Bacillus subtilis	Bacteria	Firmicutes	Bacilli
Bacillus sp YSP-3	Bacteria	Firmicutes	Bacilli
Bacillus megaterium	Bacteria	Firmicutes	Bacilli
Bacillus licheniformis	Bacteria	Firmicutes	Bacilli
Bacillus halodurans	Bacteria	Firmicutes	Bacilli
Bacillus cereus	Bacteria	Firmicutes	Bacilli
Bacillus anthracis	Bacteria	Firmicutes	Bacilli
Thermoanaerobacterium thermosacch.	Bacteria	Firmicutes	Clostridia
Pelotomaculum propionicicum	Bacteria	Firmicutes	Clostridia
Oxobacter pfennigii	Bacteria	Firmicutes	Clostridia
Hungatella hathewayi	Bacteria	Firmicutes	Clostridia
Harryflintia acetispora	Bacteria	Firmicutes	Clostridia
Fervidicola ferrireducens	Bacteria	Firmicutes	Clostridia
Desulfosporosinus meridiei	Bacteria	Firmicutes	Clostridia
Clostridium septicum	Bacteria	Firmicutes	Clostridia
Clostridium populeti	Bacteria	Firmicutes	Clostridia
Clostridium botulinum	Bacteria	Firmicutes	Clostridia
Clostridium bornimense	Bacteria	Firmicutes	Clostridia
Clostridium bolteae	Bacteria	Firmicutes	Clostridia
Clostridium aceticum	Bacteria	Firmicutes	Clostridia
Butyrivibrio proteoclasticus	Bacteria	Firmicutes	Clostridia
Anaerotignum neopropionicum	Bacteria	Firmicutes	Clostridia
Acetivibrio ethanolgignens	Bacteria	Firmicutes	Clostridia
Dubosiella newyorkensis	Bacteria	Firmicutes	Erysipelotrichia
Limnochorda pilosa	Bacteria	Firmicutes	Limnochordia
Sporomusa sphaeroides	Bacteria	Firmicutes	Negativicutes
Sporomusa malonica	Bacteria	Firmicutes	Negativicutes

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Sporomusa acidovorans	Bacteria	Firmicutes	Negativicutes
Megasphaera micronuciformis	Bacteria	Firmicutes	Negativicutes
Tissierella creatinophila	Bacteria	Firmicutes	Tissierellia
Leptotrichia buccalis	Bacteria	Fusobacteria	Fusobacteriia
Gemmatimonas phototrophica	Bacteria	Gemmatimonadetes	Gemmatimonadetes
Melioribacter roseus	Bacteria	Ignavibacteriae	Ignavibacteria
Kiritimatiella glycovorans	Bacteria	Kiritimatiellaeota	Kiritimatiellae
Nitrospira lenta	Bacteria	Nitrospirae	Nitrospira
Nitrospira defluvii	Bacteria	Nitrospirae	Nitrospira
Kuenenia stuttgartiensis	Bacteria	Planctomycetes	Cand.
Phycisphaera sp	Bacteria	Planctomycetes	Phycisphaerae
Rhodopirellula sp	Bacteria	Planctomycetes	Planctomycetia
Gemmata sp	Bacteria	Planctomycetes	Planctomycetia
Acidithiobacillus sp	Bacteria	Proteobacteria	Acidithiobacillia
Thalassobius mediterraneus	Bacteria	Proteobacteria	Alphaproteobacteria
Sedimentitalea sp	Bacteria	Proteobacteria	Alphaproteobacteria
Roseovarius albus	Bacteria	Proteobacteria	Alphaproteobacteria
Roseomonas sp	Bacteria	Proteobacteria	Alphaproteobacteria
Roseobacter denitrificans	Bacteria	Proteobacteria	Alphaproteobacteria
Roseisalinus antarcticus	Bacteria	Proteobacteria	Alphaproteobacteria
Rhodoplanes piscinae	Bacteria	Proteobacteria	Alphaproteobacteria
Pseudooceanicola marinus	Bacteria	Proteobacteria	Alphaproteobacteria
Planktotalea frisia	Bacteria	Proteobacteria	Alphaproteobacteria
Phaeobacter sp	Bacteria	Proteobacteria	Alphaproteobacteria
Phaeobacter gallaeciensis	Bacteria	Proteobacteria	Alphaproteobacteria
Neorhizobium sp	Bacteria	Proteobacteria	Alphaproteobacteria
Maliponia aquimaris	Bacteria	Proteobacteria	Alphaproteobacteria
Labrenzia sp	Bacteria	Proteobacteria	Alphaproteobacteria
Labrenzia alexandrii	Bacteria	Proteobacteria	Alphaproteobacteria
Labrenzia alba	Bacteria	Proteobacteria	Alphaproteobacteria
Bradyrhizobium diazoefficiens	Bacteria	Proteobacteria	Alphaproteobacteria
Azospirillum lipoferum	Bacteria	Proteobacteria	Alphaproteobacteria
Ascidiaceihabitans donghaensis	Bacteria	Proteobacteria	Alphaproteobacteria
Agrobacterium fabrum	Bacteria	Proteobacteria	Alphaproteobacteria
Thiomonas delicata	Bacteria	Proteobacteria	Betaproteobacteria
Ralstonia solanacearum	Bacteria	Proteobacteria	Betaproteobacteria
Pusillimonas sp	Bacteria	Proteobacteria	Betaproteobacteria
Paucimonas lemoignei	Bacteria	Proteobacteria	Betaproteobacteria
Paraburkholderia ribeironis	Bacteria	Proteobacteria	Betaproteobacteria
Pandoraea sputorum	Bacteria	Proteobacteria	Betaproteobacteria
Massilia glaciei	Bacteria	Proteobacteria	Betaproteobacteria
Limnohabitans sp	Bacteria	Proteobacteria	Betaproteobacteria
Janthinobacterium sp KBS0711	Bacteria	Proteobacteria	Betaproteobacteria
Janthinobacterium sp HH01	Bacteria	Proteobacteria	Betaproteobacteria
Duganella phyllosphaerae	Bacteria	Proteobacteria	Betaproteobacteria
Cupriavidus necator	Bacteria	Proteobacteria	Betaproteobacteria
Comamonas terrigena	Bacteria	Proteobacteria	Betaproteobacteria
Chromobacterium violaceum	Bacteria	Proteobacteria	Betaproteobacteria
Burkholderia sp	Bacteria	Proteobacteria	Betaproteobacteria
Burkholderia multivorans	Bacteria	Proteobacteria	Betaproteobacteria
Burkholderia dabaoshanensis	Bacteria	Proteobacteria	Betaproteobacteria
Bordetella hinzii	Bacteria	Proteobacteria	Betaproteobacteria
Acidovorax sp	Bacteria	Proteobacteria	Betaproteobacteria
Achromobacter piechaudii	Bacteria	Proteobacteria	Betaproteobacteria
Plesiocystis pacifica	Bacteria	Proteobacteria	Deltaproteobacteria
Myxococcus stipitatus	Bacteria	Proteobacteria	Deltaproteobacteria
Geobacter sulfurreducens	Bacteria	Proteobacteria	Deltaproteobacteria
Dethiosulfatarculus sandiegensis	Bacteria	Proteobacteria	Deltaproteobacteria
Desulfuromonas sp	Bacteria	Proteobacteria	Deltaproteobacteria
Desulfovibrio sp	Bacteria	Proteobacteria	Deltaproteobacteria
Desulfovibrio alaskensis	Bacteria	Proteobacteria	Deltaproteobacteria
Desulfofustis glycolicus	Bacteria	Proteobacteria	Deltaproteobacteria

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Desulfococcus multivorans	Bacteria	Proteobacteria	Deltaproteobacteria
Pseudoarcobacter caeni	Bacteria	Proteobacteria	Epsilonproteobacteria
Helicobacter pylori	Bacteria	Proteobacteria	Epsilonproteobacteria
Arcobacter mytili	Bacteria	Proteobacteria	Epsilonproteobacteria
Yersinia pestis	Bacteria	Proteobacteria	Gammaproteobacteria
Xenorhabdus nematophila	Bacteria	Proteobacteria	Gammaproteobacteria
Vibrio proteolyticus	Bacteria	Proteobacteria	Gammaproteobacteria
Vibrio nigripulchritudo	Bacteria	Proteobacteria	Gammaproteobacteria
Vibrio nereis	Bacteria	Proteobacteria	Gammaproteobacteria
Vibrio cholerae serotype O1	Bacteria	Proteobacteria	Gammaproteobacteria
Trabulsiella guamensis	Bacteria	Proteobacteria	Gammaproteobacteria
Tatumella ptyseos ATCC33301	Bacteria	Proteobacteria	Gammaproteobacteria
Tatumella morbirosei	Bacteria	Proteobacteria	Gammaproteobacteria
Superficieibacter electus	Bacteria	Proteobacteria	Gammaproteobacteria
Shigella flexneri	Bacteria	Proteobacteria	Gammaproteobacteria
Shigella dysenteriae serotype 1	Bacteria	Proteobacteria	Gammaproteobacteria
Shewanella oneidensis	Bacteria	Proteobacteria	Gammaproteobacteria
Salmonella typhimurium	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudoxanthomonas composti	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas syringae pv tomato	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas resinovorans	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas putida	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas fluores. F113	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas fluores. ATCCBAA477	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudomonas aeruginosa	Bacteria	Proteobacteria	Gammaproteobacteria
Pseudescherichia vulneris	Bacteria	Proteobacteria	Gammaproteobacteria
Pluralibacter gergoviae	Bacteria	Proteobacteria	Gammaproteobacteria
Photobacterium aquimaris	Bacteria	Proteobacteria	Gammaproteobacteria
Photobacterium aphoticum	Bacteria	Proteobacteria	Gammaproteobacteria
Marinomonas spartinae	Bacteria	Proteobacteria	Gammaproteobacteria
Kluyvera cryocrescens	Bacteria	Proteobacteria	Gammaproteobacteria
Klebsiella pneumoniae HS11286	Bacteria	Proteobacteria	Gammaproteobacteria
Klebsiella pneumoniae ATCC700721	Bacteria	Proteobacteria	Gammaproteobacteria
Halomonas elongata	Bacteria	Proteobacteria	Gammaproteobacteria
Haemophilus influenzae	Bacteria	Proteobacteria	Gammaproteobacteria
Franconibacter pulveris	Bacteria	Proteobacteria	Gammaproteobacteria
Escherichia coli O157:H7	Bacteria	Proteobacteria	Gammaproteobacteria
Escherichia coli K12	Bacteria	Proteobacteria	Gammaproteobacteria
Erwinia mallotivora	Bacteria	Proteobacteria	Gammaproteobacteria
Erwinia gerundensis	Bacteria	Proteobacteria	Gammaproteobacteria
Enterobacter cloacae S611	Bacteria	Proteobacteria	Gammaproteobacteria
Citrobacter koseri	Bacteria	Proteobacteria	Gammaproteobacteria
Atlantibacter hermannii	Bacteria	Proteobacteria	Gammaproteobacteria
Halobacteriovorax sp	Bacteria	Proteobacteria	Oligoflexia
Bacteriovorax stolpii	Bacteria	Proteobacteria	Oligoflexia
Mariprofundus sp	Bacteria	Proteobacteria	Zetaproteobacteria
Rubricoccus marinus	Bacteria	Rhodothermaeota	Rhodothermia
Treponema berlinense	Bacteria	Spirochaetes	Spirochaetia
Spirochaeta lutea	Bacteria	Spirochaetes	Spirochaetia
Cloacibacillus sp	Bacteria	Synergistetes	Synergistia
Mycoplasma synoviae	Bacteria	Tenericutes	Mollicutes
Mycoplasma orale	Bacteria	Tenericutes	Mollicutes
Mycoplasma marinum	Bacteria	Tenericutes	Mollicutes
Thermosulfurimonas dismutans	Bacteria	Thermodesulfobacteria	Thermodesulfobacteria
Pseudothermotoga thermarum	Bacteria	Thermotogae	Thermotogae
Mesotoga prima	Bacteria	Thermotogae	Thermotogae
Opitutus terrae	Bacteria	Verrucomicrobia	Opitutae
Opitutaceae bacterium	Bacteria	Verrucomicrobia	Opitutae
Terrimicrobium sacchariphilum	Bacteria	Verrucomicrobia	Spartobacteria
Rubritalea profundi	Bacteria	Verrucomicrobia	Verrucomicrobiae
Cand. Bathyarchaeota archaeon	Archaea	Cand. Bathyarchaeota	undef. Bathyarchaeota
Natronomonas sp	Archaea	Euryarchaeota	Halobacteria

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Natronomonas pharaonis	Archaea	Euryarchaeota	Halobacteria
Natronomonas moolapensis	Archaea	Euryarchaeota	Halobacteria
Natronolimnobius sp	Archaea	Euryarchaeota	Halobacteria
Natronolimnobius baerhuensis	Archaea	Euryarchaeota	Halobacteria
Natronococcus occultus	Archaea	Euryarchaeota	Halobacteria
Natronobacterium gregoryi	Archaea	Euryarchaeota	Halobacteria
Natronoarchaeum philippinense	Archaea	Euryarchaeota	Halobacteria
Natrinema pellirubrum	Archaea	Euryarchaeota	Halobacteria
Natrialba magadii	Archaea	Euryarchaeota	Halobacteria
Natrialba asiatica	Archaea	Euryarchaeota	Halobacteria
Natrarchaeobius chitinivorans	Archaea	Euryarchaeota	Halobacteria
Haloterrigena limicola	Archaea	Euryarchaeota	Halobacteria
Haloterrigena daqingensis	Archaea	Euryarchaeota	Halobacteria
Halostagnicola larsenii	Archaea	Euryarchaeota	Halobacteria
Halosimplex carlsbadense	Archaea	Euryarchaeota	Halobacteria
Halorubrum vacuolatum	Archaea	Euryarchaeota	Halobacteria
Halorubrum halodurans	Archaea	Euryarchaeota	Halobacteria
Halorientalis sp	Archaea	Euryarchaeota	Halobacteria
Halorhabdus tiamatea	Archaea	Euryarchaeota	Halobacteria
Haloquadratum walsbyi	Archaea	Euryarchaeota	Halobacteria
Halopiger xanaduensis	Archaea	Euryarchaeota	Halobacteria
Halopiger salifodinae	Archaea	Euryarchaeota	Halobacteria
Halophilic archaeon	Archaea	Euryarchaeota	Halobacteria
Halopenitus malekzadehii	Archaea	Euryarchaeota	Halobacteria
Halopelagius inordinatus	Archaea	Euryarchaeota	Halobacteria
Halomicrobium zhouii	Archaea	Euryarchaeota	Halobacteria
Halomicrobium mukohataei	Archaea	Euryarchaeota	Halobacteria
Halomarina oriensis	Archaea	Euryarchaeota	Halobacteria
Halolamina sp	Archaea	Euryarchaeota	Halobacteria
Halolamina pelagica	Archaea	Euryarchaeota	Halobacteria
Halohasta litchfieldiae	Archaea	Euryarchaeota	Halobacteria
Halogeometricum pallidum	Archaea	Euryarchaeota	Halobacteria
Halogeometricum borinquense	Archaea	Euryarchaeota	Halobacteria
Haloferax volcanii	Archaea	Euryarchaeota	Halobacteria
Halodesulfurarchaeum formicicum	Archaea	Euryarchaeota	Halobacteria
Halococcus saccharolyticus	Archaea	Euryarchaeota	Halobacteria
Halobiforma lacisalsi	Archaea	Euryarchaeota	Halobacteria
Halobellus limi	Archaea	Euryarchaeota	Halobacteria
Halobellus clavatus	Archaea	Euryarchaeota	Halobacteria
Halobaculum gomorrense	Archaea	Euryarchaeota	Halobacteria
Halobacterium sp	Archaea	Euryarchaeota	Halobacteria
Halobacterium salinarum	Archaea	Euryarchaeota	Halobacteria
Halobacterium jilantaiense	Archaea	Euryarchaeota	Halobacteria
Haloarcula marismortui	Archaea	Euryarchaeota	Halobacteria
Haloarchaeobius iranensis	Archaea	Euryarchaeota	Halobacteria
Halarchaeum sp	Archaea	Euryarchaeota	Halobacteria
Halarchaeum acidiphilum	Archaea	Euryarchaeota	Halobacteria
Halapricum salinum	Archaea	Euryarchaeota	Halobacteria
Halanaeroarchaeum sulfurireducens	Archaea	Euryarchaeota	Halobacteria
Halalkalicoccus jeotgali	Archaea	Euryarchaeota	Halobacteria
Methanothermobacter thermautotroph.	Archaea	Euryarchaeota	Methanobacteria
Methanosphaera stadtmanae	Archaea	Euryarchaeota	Methanobacteria
Methanobrevibacter woesei	Archaea	Euryarchaeota	Methanobacteria
Methanobrevibacter sp NOE	Archaea	Euryarchaeota	Methanobacteria
Methanobrevibacter sp AbM4	Archaea	Euryarchaeota	Methanobacteria
Methanobrevibacter filiformis	Archaea	Euryarchaeota	Methanobacteria
Methanobacterium lacus	Archaea	Euryarchaeota	Methanobacteria
Methanobacterium congolense	Archaea	Euryarchaeota	Methanobacteria
Methanothrix soehngenii	Archaea	Euryarchaeota	Methanomicrobia
Methanospirillum hungatei	Archaea	Euryarchaeota	Methanomicrobia
Methanosphaerula palustris	Archaea	Euryarchaeota	Methanomicrobia
Methanosaeta harundinacea	Archaea	Euryarchaeota	Methanomicrobia

Continued on next page

Table C.1 – continued from previous page

Species	Superkingdom	Phylum	Class
Methanoregula formicica	Archaea	Euryarchaeota	Methanomicrobia
Methanoregula boonei	Archaea	Euryarchaeota	Methanomicrobia
Methanomethylovorans hollandica	Archaea	Euryarchaeota	Methanomicrobia
Methanolobus tindarius	Archaea	Euryarchaeota	Methanomicrobia
Methanolacinia petrolearia	Archaea	Euryarchaeota	Methanomicrobia
Methanohalophilus mahii	Archaea	Euryarchaeota	Methanomicrobia
Methanohalobium evestigatum	Archaea	Euryarchaeota	Methanomicrobia
Methanofollis sp	Archaea	Euryarchaeota	Methanomicrobia
Methanoculleus taiwanensis	Archaea	Euryarchaeota	Methanomicrobia
Methanoculleus bourgensis	Archaea	Euryarchaeota	Methanomicrobia
Methanocorpusculum labreanum	Archaea	Euryarchaeota	Methanomicrobia
Methanocella paludicola	Archaea	Euryarchaeota	Methanomicrobia
Methanocella arvoryzae	Archaea	Euryarchaeota	Methanomicrobia
Cand. Syntrophoarchaeum caldarius	Archaea	Euryarchaeota	Methanomicrobia
Cand. Syntrophoarchaeum butanivorans	Archaea	Euryarchaeota	Methanomicrobia
Cand. Methanoperedens nitroreducens	Archaea	Euryarchaeota	Methanomicrobia
Methanonatronarchaeum thermophilum	Archaea	Euryarchaeota	Methanonatronarchaeia
Cand. Methanoh. thermophilum	Archaea	Euryarchaeota	Methanonatronarchaeia
Thermoplasmatales archaeon Gpl	Archaea	Euryarchaeota	Thermoplasmata
Thermoplasmatales archaeon BRNA1	Archaea	Euryarchaeota	Thermoplasmata
Picrophilus torridus	Archaea	Euryarchaeota	Thermoplasmata
Methanomethylophilus alvus	Archaea	Euryarchaeota	Thermoplasmata
Methanomassiliicoccus intestinalis	Archaea	Euryarchaeota	Thermoplasmata
Methanogenic archaeon	Archaea	Euryarchaeota	Thermoplasmata
Cand. Methanoplasma termitum	Archaea	Euryarchaeota	Thermoplasmata
Acidiplasma cupricumulans	Archaea	Euryarchaeota	Thermoplasmata
Nitrososphaera viennensis	Archaea	Thaumarchaeota	Nitrososphaeria
Cand. Nitrosocosmicus franklandus	Archaea	Thaumarchaeota	Nitrososphaeria
Cand. Nitrosopumilus salaria	Archaea	Thaumarchaeota	undef. Thaumarchaeota

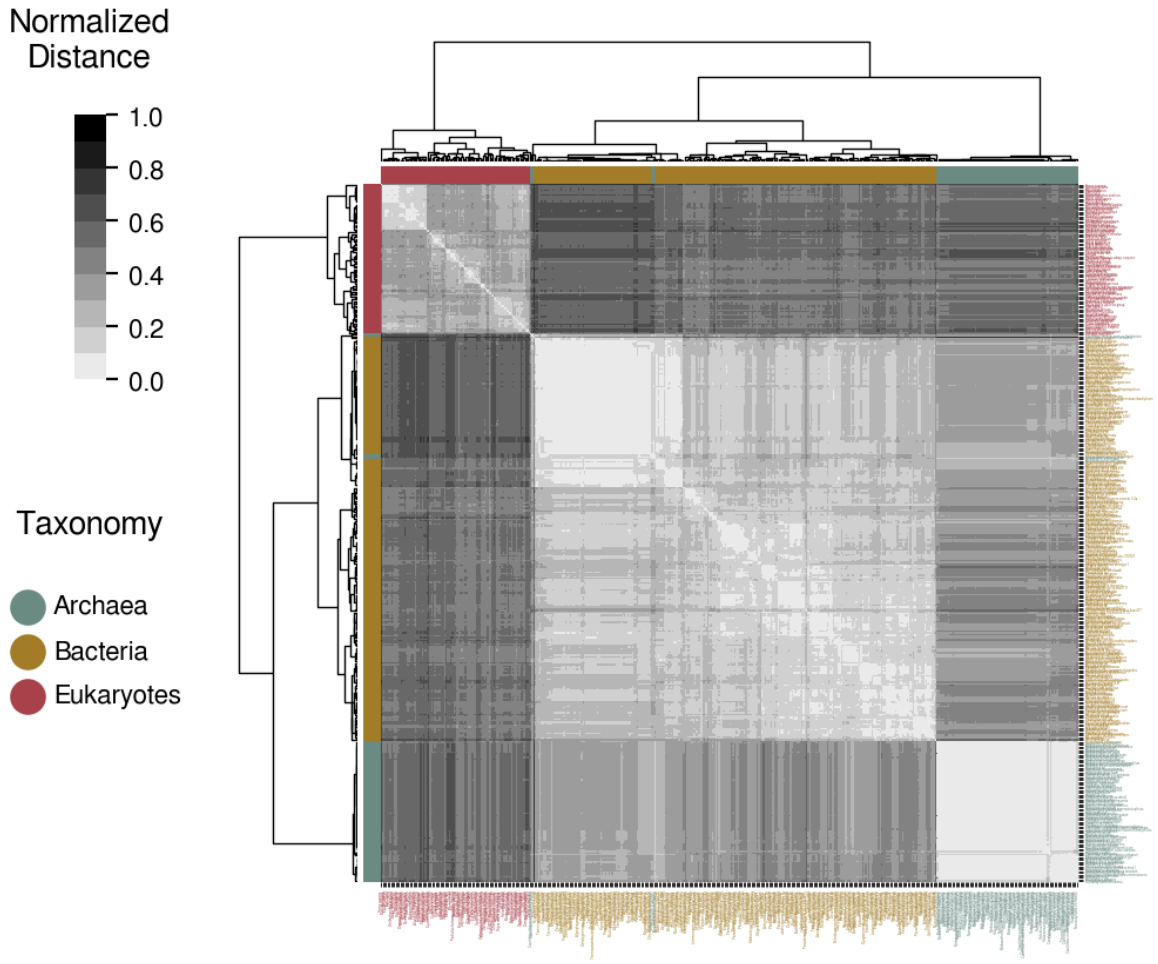


Figure C-1: Phylogenetic dendrogram based on proteostasis machinery: The tree was derived from the comparison of proteostasis functional networks (PN). These networks were constructed performing pathway analysis on proteostasis related gene sets, using the GO-BP annotation. Semantic analysis operators were applied to compare the PN networks based on their topology on the ontological graph.

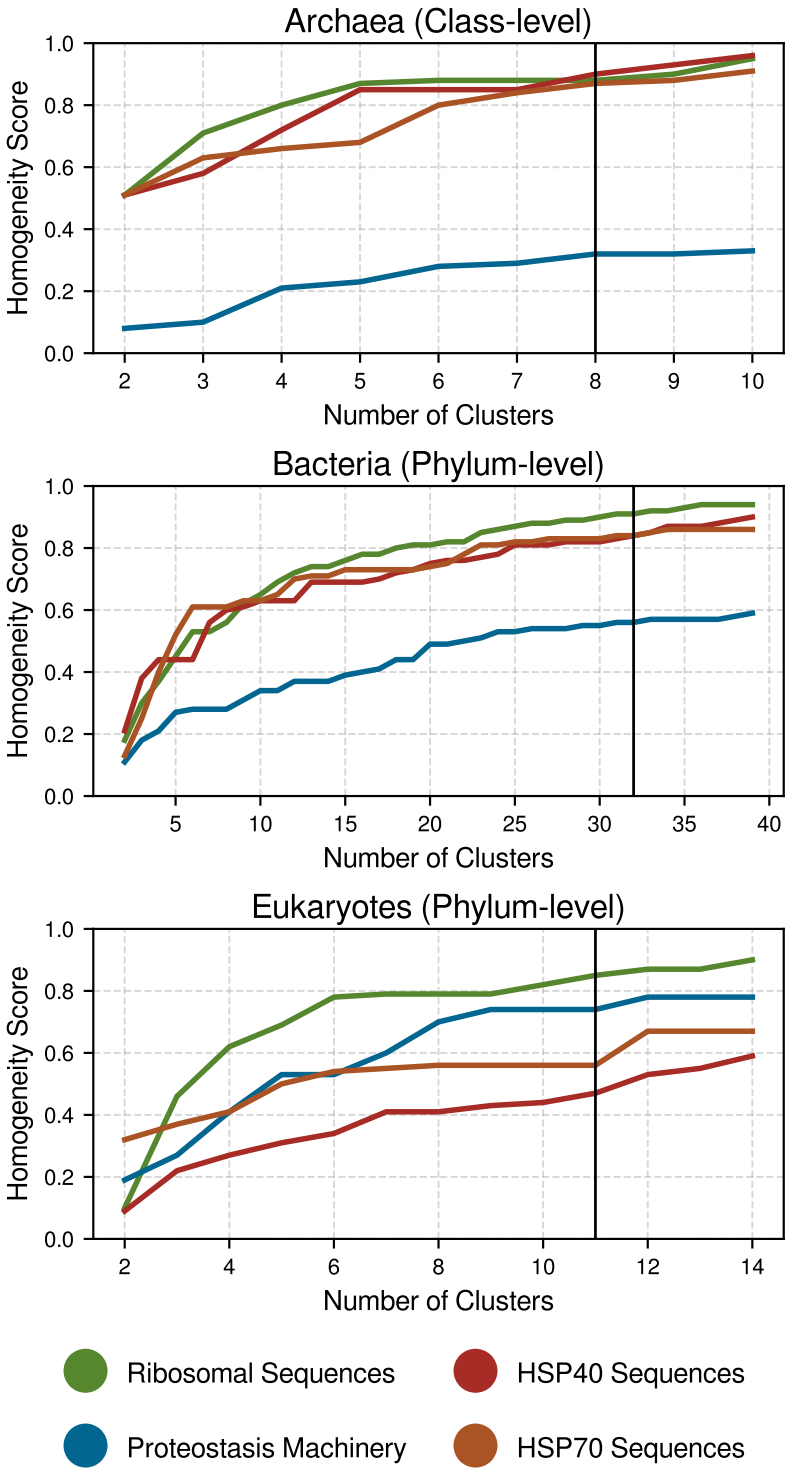


Figure C-2: Evaluation of rRNA, HSP40, HSP70 and PN to discriminate species of the same lower-level taxonomic class in Archaea, Bacteria and Eukaryotes. Different amounts of clusters were generated and for each clustering outcome the Homogeneity Score [219] was calculated, given the reference taxonomic classification of species. Vertical lines indicate the number of taxonomic groups (Phylum-level for Bacteria and Eukaryotes and Class-level for Archaea).

Appendix D

Case Study 3: Semantic Interpretation of the SARS-CoV-2 Interactome

Table D.1: Enriched GO-BP Terms of the SARS-CoV-2 Interactome

Rank	Biological Process	Enrichment	Adjusted pvalue
1	Peptide transport	73/1525	0.0007
2	Intracellular protein transport	54/976	0.0007
3	Cellular localization	99/2387	0.0019
4	Amide transport	73/1558	0.002
5	Protein transport	71/1500	0.0028
6	Intracellular transport	70/1477	0.0032
7	Establishment of protein localization to organelle	31/420	0.0043
8	Protein localization	89/2120	0.0044
9	Establishment of protein localization	72/1583	0.0055
10	Macromolecule localization	98/2442	0.0069
11	Cellular macromolecule localization	71/1570	0.0071
12	Cellular protein localization	69/1562	0.0077
13	Protein localization to organelle	42/714	0.0082
14	Protein targeting	25/351	0.0089
15	Establishment of localization in cell	75/1781	0.009
16	Cell cycle phase transition	22/268	0.0092
17	Establishment of protein localization to mitochondrion	11/72	0.0095
18	Protein targeting to mitochondrion	11/58	0.0096
19	Nitrogen compound transport	75/1820	0.0104

Continued on next page

Table D.1 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted pvalue
20	Protein localization to mitochondrion	11/79	0.0123
21	tRNA transport	8/36	0.0127
22	ncRNA export from nucleus	8/39	0.0127
23	ERAD pathway	11/84	0.0132
24	Organic substance transport	81/2183	0.0146
25	Regulation of catabolic process	44/979	0.015
26	Protein folding	17/215	0.0151
27	Mitotic cell cycle phase transition	19/260	0.0152
28	Cell cycle G2/M phase transition	13/135	0.0153
29	Protein heterotrimerization	5/13	0.0156
30	Regulation of cellular response to heat	10/79	0.0159
31	U4 snRNA 3'-end processing	4/8	0.0163
32	Viral process	33/690	0.0194
33	Protein-containing complex assembly	59/1577	0.0201
34	Organelle localization	30/564	0.0205
35	Intracellular transport of virus	8/54	0.0208
36	Mitotic cell cycle	32/684	0.021
37	Transport of virus	8/57	0.0215
38	Viral life cycle	15/198	0.0216
39	Symbiont process	35/765	0.0217
40	Protein sumoylation	8/63	0.0226
41	Endomembrane system organization	23/407	0.0227
42	Cellular component biogenesis	89/2730	0.0228
43	Nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'	4/10	0.0232
44	Chaperone-mediated protein transport	4/11	0.0235
45	mRNA transport	12/147	0.0242
46	Microtubule nucleation	5/20	0.0251
47	Mitotic cell cycle process	28/587	0.0274
48	G2/M transition of mitotic cell cycle	11/133	0.0284
49	Interspecies interaction between organisms	35/808	0.029
50	Protein-containing complex subunit organization	65/1858	0.0298
51	Membrane docking	13/178	0.0298
52	Centriole-centriole cohesion	4/12	0.0302
53	Rab protein signal transduction	8/71	0.0302
54	Golgi organization	11/132	0.0303
55	Nuclear polyadenylation-dependent tRNA catabolic process	3/5	0.0307

Continued on next page

Table D.1 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted pvalue
56	Telomere maintenance via semi-conservative replication	5/26	0.0318
57	Low-density lipoprotein particle clearance	5/24	0.0319
58	Response to endoplasmic reticulum stress	15/243	0.0343
59	Ciliary basal body-plasma membrane docking	9/95	0.0344
60	Regulation of bone development	5/25	0.0347
61	snRNA metabolic process	6/40	0.0348
62	Organelle localization by membrane tethering	12/169	0.0364
63	RNA localization	14/210	0.0371
64	Regulation of generation of precursor metabolites and energy	11/146	0.0372
65	Mitochondrion organization	22/442	0.0377
66	Regulation of glycolytic process	8/76	0.0379
67	Ribonucleoprotein complex export from nucleus	10/126	0.0384
68	RNA surveillance	4/16	0.0385
69	Protein export from nucleus	11/149	0.0397
70	Glycoprotein metabolic process	19/367	0.0408
71	Regulation of intracellular protein transport	14/230	0.0411
72	Regulation of carbohydrate catabolic process	8/84	0.0418
73	Regulation of G2/M transition of mitotic cell cycle	13/200	0.0425
74	Organelle organization	102/3363	0.0436
75	Cellular component organization or biogenesis	161/5779	0.0436
76	Ribonucleoprotein complex localization	10/127	0.0436
77	RNA export from nucleus	10/129	0.0437
78	Microtubule polymerization	5/31	0.0443
79	Nuclear export	11/158	0.0446
80	Microtubule organizing center organization	8/91	0.0458
81	Regulation of mRNA stability	12/178	0.0459
82	Regulation of RNA stability	12/184	0.0468
83	rRNA 3'-end processing	3/9	0.0472
84	Regulation of intracellular transport	18/344	0.0477
85	Ubiquitin-dependent ERAD pathway	7/66	0.0484
86	Establishment of RNA localization	12/189	0.049

Table D.2: Enriched Reactome Terms of the SARS-CoV-2 Interactome

Rank	Biological Process	Enrichment	Adjusted pvalue
1	Mitochondrial protein import	12/65	0.0015
2	Nuclear Pore Complex (NPC) Disassembly	9/36	0.0044

Continued on next page

Table D.2 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted pvalue
3	Export of Viral Ribonucleoproteins from Nucleus	8/33	0.0077
4	Viral Messenger RNA Synthesis	9/44	0.01
5	Transport of the SLBP Dependant Mature mRNA	8/36	0.0105
6	Protein localization	17/164	0.0107
7	SUMOylation of ubiquitinylation proteins	8/39	0.0149
8	M Phase	28/394	0.017
9	Nuclear Envelope Breakdown	9/53	0.019
10	NS1 Mediated Effects on Host Pathways	8/41	0.0212
11	Transport of Mature mRNA Derived from an Intronless Transcript	8/42	0.0217
12	Cell Cycle, Mitotic	34/538	0.0227
13	SUMOylation of DNA replication proteins	8/46	0.0274
14	Host Interactions of HIV factors	14/131	0.029
15	SUMOylation of RNA binding proteins	8/47	0.0308
16	Recruitment of NuMA to mitotic centrosomes	11/94	0.0329
17	SUMOylation of chromatin organization proteins	9/71	0.0371
18	Cell Cycle	36/644	0.0374
19	Regulation of PLK1 Activity at G2/M Transition	10/87	0.0382
20	Mitotic Prophase	13/143	0.0431
21	tRNA processing in the nucleus	8/59	0.0442
22	ISG15 antiviral mechanism	9/74	0.0456
23	Transport of Mature mRNA derived from an Intron-Containing Transcript	9/75	0.0484
24	Anchoring of the basal body to the plasma membrane	10/97	0.0496

Table D.3: Enriched MGIMP Terms of the SARS-CoV-2 Interactome

Rank	Biological Process	Enrichment	Adjusted pvalue
1	Preweaning lethality	125/4062	0.0011
2	Abnormal survival	131/4494	0.0019
3	Mortality/aging	133/4820	0.0033
4	Preweaning lethality, complete penetrance	47/1316	0.0035
5	Abnormal ascending aorta morphology	4/18	0.0044
6	Abnormal aorta morphology	12/197	0.006
7	Accumulation of giant lysosomes in kidney/renal tubule cells	3/8	0.0061
8	Abnormal aorta elastic tissue morphology	4/20	0.0063
9	Abnormal blood vessel elastic tissue morphology	4/22	0.0063
10	Loss of cortex neurons	3/11	0.0092

Continued on next page

Table D.3 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted pvalue
11	Perinatal lethality	33/985	0.0103
12	Abnormal thoracic aorta morphology	8/117	0.0118
13	Abnormal aorta wall morphology	5/45	0.012
14	Increased total body fat amount	11/202	0.012
15	Abnormal aorta elastic fiber morphology	3/13	0.0128
16	Abnormal heart right ventricle morphology	9/148	0.0137
17	Abnormal active avoidance behavior	3/15	0.0151
18	Abnormal aorta tunica media morphology	4/35	0.016
19	Abnormal blastocoele morphology	4/32	0.0166
20	Abnormal embryo development	31/977	0.0172
21	Embryonic lethality prior to tooth bud stage	28/842	0.0174
22	Prenatal lethality	58/2126	0.0174
23	Abnormal basal metabolism	4/37	0.0205
24	Abnormal systemic artery morphology	12/261	0.0213
25	Embryonic lethality prior to organogenesis	24/748	0.0216
26	Abnormal pulmonary circulation	5/61	0.0223
27	Abnormal digit development	3/21	0.0223
28	Abnormal vacuole morphology	4/45	0.0237
29	Abnormal lysosome morphology	4/44	0.0251
30	Lethality during fetal growth through weaning	47/1765	0.0251
31	Embryonic lethality	45/1643	0.0252
32	Abnormal artery morphology	15/398	0.0255
33	Absent pinna reflex	5/70	0.0278
34	Decreased vertical activity	8/160	0.0302
35	Internal hemorrhage	11/272	0.0319
36	Abnormal thrombosis	5/75	0.0323
37	Embryo phenotype	45/1689	0.0331
38	Abnormal RR interval	4/49	0.0338
39	Abnormal vertical activity	10/241	0.034
40	Absent blastocoele	3/29	0.0345
41	Abnormal developmental patterning	15/429	0.0364
42	Perinatal lethality, incomplete penetrance	10/237	0.0366
43	Abnormal blood coagulation	9/209	0.0376
44	Abnormal hemostasis	9/211	0.0385
45	Abnormal pinna reflex	5/80	0.0389
46	Preweaning lethality, incomplete penetrance	20/630	0.0403
47	Abnormal brain size	10/249	0.0405
48	Abnormal adipose tissue amount	24/806	0.0416
49	Abnormal embryo implantation	4/55	0.0417

Continued on next page

Table D.3 – continued from previous page

Rank	Biological Process	Enrichment	Adjusted pvalue
50	Edema	15/439	0.0423
51	Lethality throughout fetal growth and development, incomplete penetrance	8/183	0.0429
52	Hemorrhage	16/491	0.0433
53	Neurodegeneration	13/372	0.045
54	Abnormal circulating amylase level	4/61	0.0451
55	Coloboma	3/33	0.0465
56	Nervous system inclusion bodies	3/34	0.047
57	Abnormal embryonic neuroepithelium morphology	5/91	0.0474
58	Abnormal red blood cell distribution width	6/124	0.0498
59	Abnormal heart ventricle morphology	18/589	0.0499

Appendix E

Publications

Journal Publications

- **T. Koutsandreas**, E. Chevet, A. Chatziioannou, and B. Felden, "Protein homeostasis imprinting across evolution," *bioRxiv*, Jul. 2020. DOI:10.1101/2020.07.19.210591

- D. Sicari, A. Chatziioannou, **T. Koutsandreas**, R. Sitia, and E. Chevet, "Role of the early secretory pathway in SARS-CoV-2 infection," *The Journal of Cell Biology*, vol. 219, no.9, Nov. 2020. DOI:10.1083/jcb.202006005

- L. Cardoso Alves, M. D. Berger, **T. Koutsandreas**, N. Kirschke, C. Lauer, R. Spörri, A. Chatziioannou, N. Corazza, and P. Krebs, "Non-apoptotic TRAIL function modulates NK cell activity during viral infection," *EMBO reports*, vol.21, no.1, e48789–e48789, Jan. 2020. DOI:10.15252/embr.201948789

- I. Binenbaum, H. Abu-Toamih Atamni, G. Fotakis, G. Kontogianni, **T. Koutsandreas**, E. Pilalis, R. Mott, H. Himmelbauer, F. Iraqi, and A. Chatziioannou, "Container-aided integrative QTL and RNA-seq analysis of Collaborative Cross mice supports distinct sex-oriented molecular modes of response in obesity," *BMC Genomics*, vol. 1, pp. 761–774, Nov. 2020. DOI:10.1186/s12864-020-07173-x

- **T. Koutsandreas***, E. Ladoukakis*, E. Pilalis, D. Zarafeta, F. N. Kolisis, G. Skretas, and A. A. Chatziioannou, "ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next

Generation Sequencing Data," *Frontiers in Genetics*, vol. 10, p.469, 2019. DOI:10.3389/fgene.2019.00469 (*Equal Contributors)

- **T. Koutsandreas**, I. Binenbaum, E. Pilalis, I. Valavanis, O. Papadodima, and A. Chatziioannou, "Analyzing and Visualizing Genomic Complexity for the Derivation of the Emergent Molecular Networks," *Int. J. Monit. Surveill. Technol. Res.*, vol.4, no.2, pp.30–49, Apr. 2016. DOI:10.4018/IJMSTR.2016040103

- E. Pilalis, **T. Koutsandreas**, I. Valavanis, E. Athanasiadis, G. Spyrou, and A. Chatziioannou, "KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 248–255, 2015. DOI:10.1016/j.csbj.2015.03.009

Conference Publications/Presentations

- **T. Koutsandreas**, A. Bajram, C. Mastrokalou, E. Pilalis, A. Chatziioannou, and I. Maglogiannis, "Combining pathway analysis and supervised machine learning for the functional classification of single-cell transcriptomic data," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct. 2019, pp. 861–866. DOI:10.1109/BIBE.2019.00160

- A. Xenos, **T. Koutsandreas**, and A. Chatziioannou, "Proposing a unified framework of topological factors in order to refine semantic network analysis on biomedical ontologies," *Talk at Bio-Ontologies COSI Track ISMB/ECCB 2019, Basel, Switzerland*

- E. Pilalis, **T. Koutsandreas**, and A. Chatziioannou, "Streamlined, unsupervised biomarker signature derivation with the e-NIOS BioInfoMiner2.0 integrative -omics data analysis and interpretation platform," *Talk at 17th European Conference on Computational Biology (2018), Athens, Greece*

- **T. Koutsandreas** and A. Chatziioannou, "Automated in-silico characterization of the enzymatic universe,through machine-learning, protein domain-based,analysis," *Talk at 12th Conference of the Hellenic Society for Computational Biology and Bioinformatics (2017), Athens, Greece*

- **T. Koutsandreas**, E. Pilalis, E. I. Vlachavas, D. Koczan, S. Klippel, A. Dimitrakopoulou-Strauss, I. Valavanis, and A. Chatziioannou, "Making

sense of the biological complexity through the platform-driven unification of the analytical and visualization tasks," in *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2015, pp. 1–6. DOI:10.1109/BIBE.2015.7367724

- **T. Koutsandreas**, I. Valavanis, E. Pilalis, and A. Chatziioannou, "An entropy-based statistical workflow provides noise-minimizing biological annotation for muscular aging," in *2014 8th International Conference on Systems Biology (ISB)*, Oct. 2014, pp. 156–163. DOI:10.1109/ISB.2014.6990749

Bibliography

[1] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009. DOI: 10.1126/science.1170411.

[2] T. Hey, D. Gannon, and J. Pinkelman, "The Future of Data-Intensive Science," *Computer*, vol. 45, no. 5, pp. 81–82, 2012. DOI: 10.1109/MC.2012.181.

[3] B. Smith and B. Klagges, "Philosophy and Biomedical Information Systems," in *Applied Ontology*. Berlin, Boston: De Gruyter, 2008, pp. 21–38. DOI: 10.1515/9783110324860.21.

[4] A. Thomasson, "Categories," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Summer 2019, Metaphysics Research Lab, Stanford University, 2019.

[5] A. C. Yu, "Methods in biomedical ontology," *Journal of Biomedical Informatics*, vol. 39, no. 3, pp. 252–266, 2006, Biomedical Ontologies. DOI: 10.1016/j.jbi.2005.11.006.

[6] D. Rubin, N. Shah, and N. Noy, "Biomedical ontologies: A functional perspective," *Briefings in bioinformatics*, vol. 9, pp. 75–90, Feb. 2008. DOI: 10.1093/bib/bbm059.

[7] J. Hastings, "Primer on Ontologies," in *The Gene Ontology Handbook*, C. Dessimoz and N. Škunca, Eds. New York, NY: Springer New York, 2017, pp. 3–13. DOI: 10.1007/978-1-4939-3743-1_1.

[8] T. G. O. Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, Nov. 2018. DOI: 10.1093/nar/gky1055.

[9] E. H. Granger, "Aristotle on Genus and Differentia," *Journal of the History of Philosophy*, vol. 22, no. 1, pp. 1–23, 1984. DOI: 10.1353/hph.1984.0001.

[10] B. Smith, W. Ceusters, B. Klagges, *et al.*, "Relations in Biomedical Ontologies," *Genome biology*, vol. 6, R46, Feb. 2005. DOI: 10.1186/gb-2005-6-5-r46.

[11] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Briefings in Bioinformatics*, vol. 16, pp. 1069–1080, 2015. DOI: 10.1093/bib/bbv011.

[12] V. McKusick, "Mendelian Inheritance in Man and Its Online Version, OMIM," *American journal of human genetics*, vol. 80, pp. 588–604, May 2007. DOI: 10.1086/514346.

[13] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, Jan. 2000. DOI: 10.1093/nar/28.1.27.

[14] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Science*, vol. 28, no. 11, pp. 1947–1951, 2019. DOI: 10.1002/pro.3715.

[15] P. D. Karp, M. Riley, S. M. Paley, and A. Pelligrini-Toole, "EcoCyc: An Encyclopedia of Escherichia Coli Genes and Metabolism," *Nucleic Acids Research*, vol. 24, no. 1, pp. 32–39, Jan. 1996. DOI: 10.1093/nar/24.1.32.

[16] I. M. Keseler, A. Mackie, A. Santos-Zavaleta, *et al.*, "The EcoCyc database: reflecting new knowledge about Escherichia coli K-12," *Nucleic Acids Research*, vol. 45, no. D1, pp. D543–D550, Nov. 2016. DOI: 10.1093/nar/gkw1003.

[17] M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000. DOI: 10.1038/75556.

[18] J. C. Venter, M. D. Adams, E. W. Myers, *et al.*, "The Sequence of the Human Genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001. DOI: 10.1126/science.1058040.

[19] T. M. Powledge, "Human genome project completed," *Genome Biology*, vol. 4, no. 1, Apr. 2003. DOI: 10.1186/gb-spotlight-20030415-01.

[20] A. J. Iafrate, L. Feuk, M. N. Rivera, *et al.*, "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, no. 9, pp. 949–951, Sep. 2004. DOI: 10.1038/ng1416.

[21] J. Sebat, B. Lakshmi, J. Troge, *et al.*, "Large-Scale Copy Number Polymorphism in the Human Genome," *Science*, vol. 305, no. 5683, pp. 525–528, 2004. DOI: 10.1126/science.1098918.

[22] D. G. Albertson and D. Pinkel, "Genomic microarrays in human genetic disease and cancer," *Human Molecular Genetics*, vol. 12, R145–R152, Oct. 2003. DOI: 10.1093/hmg/ddg261.

[23] D. Gresham, M. J. Dunham, and D. Botstein, "Comparing whole genomes using DNA microarrays," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 291–302, Apr. 2008. DOI: 10.1038/nrg2335.

[24] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, Jun. 2016. DOI: 10.1038/nrg.2016.49.

[25] M. Bada, R. Stevens, C. Goble, *et al.*, "A short study on the success of the Gene Ontology," *Journal of Web Semantics*, vol. 1, no. 2, pp. 235–240, 2004, 2003 World Wide Web Conference. DOI: 10.1016/j.websem.2003.12.003.

[26] B. Jassal, L. Matthews, G. Viteri, *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 48, no. D1, pp. D498–D503, Nov. 2019. DOI: 10.1093/nar/gkz1031.

[27] A. Morgat, E. Coissac, E. Coudert, *et al.*, "UniPathway: a resource for the exploration and annotation of metabolic pathways," *Nucleic acids research*, vol. 40, no. Database issue, Jan. 2012. DOI: 10.1093/nar/gkr1023.

[28] L. M. Schriml, E. Mitraka, J. Munro, *et al.*, "Human Disease Ontology 2018 update: classification, content and workflow expansion," *Nucleic acids research*, vol. 47, no. D1, pp. D955–D962, Jan. 2019. DOI: 10.1093/nar/gky1032.

[29] S. Köhler, L. Carmody, N. Vasilevsky, *et al.*, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1018–D1027, Nov. 2018. DOI: 10.1093/nar/gky1105.

[30] C. L. Smith and J. T. Eppig, "The mammalian phenotype ontology: enabling robust annotation and comparative analysis," *WIREs Systems Biology and Medicine*, vol. 1, no. 3, pp. 390–399, 2009. DOI: 10.1002/wsbm.44.

[31] M. A. Harris, A. Lock, J. Bähler, S. G. Oliver, and V. Wood, "FYPO: the fission yeast phenotype ontology," *Bioinformatics*, vol. 29, no. 13, pp. 1671–1678, Jul. 2013. DOI: 10.1093/bioinformatics/btt266.

[32] E. Arnaud, L. Cooper, N. Menda, *et al.*, "Towards a Reference Plant Trait Ontology For Modeling Knowledge of Plant Traits and Phenotypes," Oct. 2012. DOI: 10.13140/2.1.2550.3525.

[33] T. F. Hayamizu, R. A. Baldock, and M. Ringwald, "Mouse anatomy ontologies: enhancements and tools for exploring and integrating biomedical data," *Mammalian Genome*, vol. 26, no. 9, pp. 422–430, Oct. 2015. DOI: 10.1007/s00335-015-9584-9.

[34] E. Segerdell, J. B. Bowes, N. Pollet, and P. D. Vize, "An ontology for Xenopus anatomy and development," *BMC Developmental Biology*, vol. 8, no. 1, p. 92, Sep. 2008. DOI: 10.1186/1471-213X-8-92.

[35] C. E. Van Slyke, Y. M. Bradford, M. Westerfield, and M. A. Haendel, "The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio," *Journal of Biomedical Semantics*, vol. 5, no. 1, p. 12, Feb. 2014. DOI: 10.1186/2041-1480-5-12.

[36] T. W. Harris, V. Arnaboldi, S. Cain, *et al.*, "WormBase: a modern Model Organism Information Resource," *Nucleic Acids Research*, vol. 48, no. D1, pp. D762–D767, Oct. 2019. DOI: 10.1093/nar/gkz920.

[37] J. R. Smith, C. A. Park, R. Nigam, *et al.*, "The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 26, Oct. 2013. DOI: 10.1186/2041-1480-4-26.

[38] G. Mayer, L. Montecchi-Palazzi, D. Ovelleiro, *et al.*, "The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary," *Database : the journal of biological databases and curation*, vol. 2013, bat009–bat009, Mar. 2013. DOI: 10.1093/database/bat009.

[39] A. D. Diehl, T. F. Meehan, Y. M. Bradford, *et al.*, "The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability," *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 44, Jul. 2016. DOI: 10.1186/s13326-016-0088-7.

[40] W. O. W. Group, *OWL 2 Web Ontology Language Document Overview (Second Edition)*, https://www.w3.org/TR/2012/REC-owl2-overview-20121211/, W3C, Dec. 2012.

[41] B. Smith, M. Ashburner, C. Rosse, *et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007. DOI: 10.1038/nbt1346.

[42] A. Morgat, T. Lombardot, K. B. Axelsen, *et al.*, "Updates in Rhea - an expert curated resource of biochemical reactions," *Nucleic acids research*, vol. 45, no. D1, pp. D415–D418, Jan. 2017. DOI: 10.1093/nar/gkw990.

[43] S. Orchard, M. Ammari, B. Aranda, *et al.*, "The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, vol. 42, no. Database issue, pp. D358–63, Jan. 2014. DOI: 10.1093/nar/gkt1115.

[44] R. Caspi, T. Altman, R. Billington, *et al.*, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Research*, vol. 42, no. D1, pp. D459–D471, Nov. 2013. DOI: 10.1093/nar/gkt1103.

[45] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, Jan. 2004. DOI: 10.1093/nar/gkh061.

[46] I. K. Dhammi and S. Kumar, "Medical subject headings (MeSH) terms," *Indian journal of orthopaedics*, vol. 48, no. 5, pp. 443–444, Sep. 2014. DOI: 10.4103/0019-5413.139827.

[47] *Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net.*

[48] H. V. Firth, S. M. Richards, A. P. Bevan, *et al.*, "DECIPHER: *D*atabase of *C*hromosomal *I*mbalance and *P*henotype in *H*umans Using *E*nsembl *R*esources," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 524–533, Apr. 2009. DOI: 10.1016/j.ajhg.2009.03.010.

[49] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, "Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 449–462, Aug. 2011. DOI: 10.1093/bib/bbr042.

[50] D. P. Hill, B. Smith, M. S. McAndrews-Hill, and J. A. Blake, "Gene Ontology annotations: what they mean and where they come from," *BMC Bioinformatics*, vol. 9, no. 5, S2, Apr. 2008. DOI: 10.1186/1471-2105-9-S5-S2.

[51] G. Valentini, "True Path Rule Hierarchical Ensembles for Genome-Wide Gene Function Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 832–847, 2011. DOI: 10.1109/TCBB.2010.38.

[52] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-throughput sequencing technologies," *Molecular cell*, vol. 58, no. 4, pp. 586–597, May 2015. DOI: 10.1016/j.molcel.2015.05.004.

[53] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, "Pathway Analysis: State of the Art," *Frontiers in Physiology*, vol. 6, p. 383, 2015. DOI: 10.3389/fphys.2015.00383.

[54] P. Khatri, M. Sirota, and A. J. Butte, "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges," *PLOS Computational Biology*, vol. 8, no. 2, pp. 1–10, Feb. 2012. DOI: 10.1371/journal.pcbi.1002375.

[55] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, Feb. 2007. DOI: 10.1093/bioinformatics/btm051.

[56] J. J. Goeman and A. Solari, "Multiple hypothesis testing in genomics," *Statistics in Medicine*, vol. 33, no. 11, pp. 1946–1978, 2014. DOI: 10.1002/sim.6082.

[57] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[58] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinformatics*, vol. 20, no. 4, pp. 578–580, Jan. 2004. DOI: 10.1093/bioinformatics/btg455.

[59] T. Beißbarth and T. P. Speed, "GOstat: find statistically overrepresented Gene Ontologies within a group of genes," *Bioinformatics*, vol. 20, no. 9, pp. 1464–1465, Feb. 2004. DOI: 10.1093/bioinformatics/bth088.

[60] Q. Zheng and X.-J. Wang, "GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis," *Nucleic Acids Research*, vol. 36, W358–W363, May 2008. DOI: 10.1093/nar/gkn276.

[61] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks," *Bioinformatics*, vol. 21, no. 16, pp. 3448–3449, Jun. 2005. DOI: 10.1093/bioinformatics/bti551.

[62] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, no. 1, p. 48, Feb. 2009. DOI: 10.1186/1471-2105-10-48.

[63] G. Bindea, B. Mlecnik, H. Hackl, *et al.*, "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, Feb. 2009. DOI: 10.1093/bioinformatics/btp101.

[64] E. Y. Chen, C. M. Tan, Y. Kou, and et al., "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, no. 1, p. 128, Apr. 2013. DOI: 10.1186/1471-2105-14-128.

[65] M. Pomaznoy, B. Ha, and B. Peters, "GOnet: a tool for interactive Gene Ontology analysis," *BMC Bioinformatics*, vol. 19, no. 1, p. 470, Dec. 2018. DOI: 10.1186/s12859-018-2533-3.

[66] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, and J. Vilo, "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)," *Nucleic Acids Research*, vol. 47, no. W1, W191–W198, May 2019. DOI: 10.1093/nar/gkz369.

[67] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron, "Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis," *Bioinformatics*, vol. 23, no. 22, pp. 3024–3031, Sep. 2007. DOI: 10.1093/bioinformatics/btm440.

[68] A. Alexa, J. Rahnenführer, and T. Lengauer, "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure," *Bioinformatics*, vol. 22, no. 13, pp. 1600–1607, Apr. 2006. DOI: 10.1093/bioinformatics/btl140.

[69] Y. Lu, R. Rosenfeld, I. Simon, G. Nau, and Z. Bar-Joseph, "A probabilistic generative model for GO enrichment Analysis," *Nucleic acids research*, vol. 36, e109, Sep. 2008. DOI: 10.1093/nar/gkn434.

[70] S. Bauer, J. Gagneur, and P. N. Robinson, "GOing Bayesian: model-based gene set analysis of genome-scale data," *Nucleic acids research*, vol. 38, no. 11, pp. 3523–3532, Jun. 2010. DOI: 10.1093/nar/gkq045.

[71] S. Zhang, J. Cao, Y. M. Kong, and R. H. Scheuermann, "GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach," *Bioinformatics*, vol. 26, no. 7, pp. 905–911, Feb. 2010. DOI: 10.1093/bioinformatics/btq059.

[72] A. Subramanian, P. Tamayo, V. K. Mootha, *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," vol. 102, no. 43, pp. 15 545–15 550, 2005. DOI: 10.1073/pnas.0506580102.

[73] S.-Y. Kim and D. J. Volsky, "PAGE: Parametric Analysis of Gene Set Enrichment," *BMC Bioinformatics*, vol. 6, no. 1, p. 144, Jun. 2005. DOI: 10.1186/1471-2105-6-144.

[74] Y. Lu, P.-Y. Liu, P. Xiao, and H.-W. Deng, "Hotelling's T-2 multivariate profiling for detecting differential expression in microarrays," *Bioinformatics*, vol. 21, pp. 3105–13, Aug. 2005. DOI: 10.1093/bioinformatics/bti496.

[75] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," vol. 102, no. 38, pp. 13 544–13 549, 2005. DOI: 10.1073/pnas.0506577102.

[76] S. W. Kong, W. T. Pu, and P. J. Park, "A multivariate approach for integrating genome-wide expression data and biological knowledge," *Bioinformatics*, vol. 22, no. 19, pp. 2373–2380, Jul. 2006. DOI: 10.1093/bioinformatics/btl401.

[77] F. Al-Shahrour, L. Arbiza, H. Dopazo, J. Huerta-Cepas, P. Mínguez, D. Montaner, and J. Dopazo, "From genes to functional classes in the study of biological systems," *BMC bioinformatics*, vol. 8, pp. 114–114, Apr. 2007. DOI: 10.1186/1471-2105-8-114.

[78] I. Dinu, J. D. Potter, T. Mueller, *et al.*, "Improving gene set analysis of microarray data by SAM-GS," *BMC Bioinformatics*, vol. 8, no. 1, p. 242, Jul. 2007. DOI: 10.1186/1471-2105-8-242.

[79] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *Ann. Appl. Stat.*, vol. 1, no. 1, pp. 107–129, Jun. 2007. DOI: 10.1214/07-AOAS101. [Online]. Available: https://doi.org/10.1214/07-AOAS101.

[80] Z. Jiang and R. Gentleman, "Extensions to gene set enrichment," *Bioinformatics*, vol. 23, no. 3, pp. 306–313, Nov. 2006. DOI: 10.1093/bioinformatics/btl599.

[81] R. A. Irizarry, C. Wang, Y. Zhou, and T. P. Speed, "Gene set enrichment analysis made simple," *Statistical methods in medical research*, vol. 18, no. 6, pp. 565–575, Dec. 2009. DOI: 10.1177/0962280209351908.

[82] A. L. Tarca, S. Draghici, P. Khatri, *et al.*, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, Nov. 2008. DOI: 10.1093/bioinformatics/btn577.

[83] S. Gao and X. Wang, "TAPPA: topological analysis of pathway phenotype association," *Bioinformatics*, vol. 23, no. 22, pp. 3100–3102, Sep. 2007. DOI: 10.1093/bioinformatics/btm460.

[84] M. S. Massa, M. Chiogna, and C. Romualdi, "Gene set analysis exploiting the topology of a pathway," *BMC Systems Biology*, vol. 4, no. 1, p. 121, Nov. 2010. DOI: 10.1186/1752-0509-4-121.

[85] L. Jacob, P. Neuvial, and S. Dudoit, "More power via graph-structured tests for differential expression of gene networks," *The Annals of Applied Statistics*, vol. 6, no. 2, Jun. 2012, ISSN: 1932-6157. DOI: 10.1214/11-aoas528. [Online]. Available: http://dx.doi.org/10.1214/11-AOAS528.

[86] P. Martini, G. Sales, M. S. Massa, M. Chiogna, and C. Romualdi, "Along signal paths: an empirical gene set approach exploiting pathway topology," *Nucleic Acids Research*, vol. 41, no. 1, e19–e19, Sep. 2012. DOI: 10.1093/nar/gks866.

[87] B. Dutta, A. Wallqvist, and J. Reifman, "PathNet: a tool for pathway analysis using topological information," *Source Code for Biology and Medicine*, vol. 7, no. 1, p. 10, Nov. 2012. DOI: 10.1186/1751-0473-7-10.

[88] A. Alexeyenko, W. Lee, M. Pernemalm, *et al.*, "Network enrichment analysis: Extension of gene-set enrichment analysis to gene networks," *BMC Bioinformatics*, vol. 13, no. 1, p. 226, Nov. 2012. DOI: 10.1186/1471-2105-13-226.

[89] Z. Gu and J. Wang, "CePa: an R package for finding significant pathways weighted by multiple network centralities," *Bioinformatics*, vol. 29, no. 5, pp. 658–660, Jan. 2013. DOI: 10.1093/bioinformatics/btt008.

[90] D. Nishimura, "BioCarta," *Biotech Software & Internet Report*, vol. 2, no. 3, pp. 117–120, 2001. DOI: 10.1089/152791601750294344.

[91] E. Demir, M. P. Cary, S. Paley, *et al.*, "The BioPAX community standard for pathway data sharing," *Nature biotechnology*, vol. 28, no. 9, pp. 935–942, Sep. 2010, nbt.1666[PII], ISSN: 1546-1696. DOI: 10.1038/nbt.1666.

[92] M. Hucka, A. Finney, H. M. Sauro, *et al.*, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, Mar. 2003, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg015.

[93] K. Laboratories, *Kegg markup language*, 2016.

[94] D. Szklarczyk, A. L. Gable, D. Lyon, *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, Nov. 2018. DOI: 10.1093/nar/gky1131.

[95] R. Oughtred, J. Rust, C. Chang, *et al.*, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein science*, vol. 30, no. 1, pp. 187–200, Jan. 2021. DOI: 10.1002/pro.3978.

[96] A. Chatziioannou and P. Moulos, "Exploiting statistical methodologies and controlled vocabularies for prioritized functional analysis of genomic experiments: The StRAnGER web application," *Frontiers in Neuroscience*, vol. 5, p. 8, 2011. DOI: 10.3389/fnins.2011.00008.

[97] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, Jan. 1979. DOI: 10.1214/aos/1176344552.

[98] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, Nov. 2005. DOI: 10.1242/jcs.02714.

[99] P. D. Stenson, E. V. Ball, M. Mort, *et al.*, "Human Gene Mutation Database (HGMD®): 2003 update," *Human Mutation*, vol. 21, no. 6, pp. 577–581, 2003. DOI: https://doi.org/10.1002/humu.10212.

[100] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The Genetic Association Database," *Nature Genetics*, vol. 36, no. 5, pp. 431–432, May 2004. DOI: 10.1038/ng0504-431.

[101] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research*, vol. 37, W305–W311, May 2009, ISSN: 0305-1048. DOI: 10.1093/nar/gkp427.

[102] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *American journal of human genetics*, vol. 82, no. 4, pp. 949–958, Apr. 2008. DOI: 10.1016/j.ajhg.2008.02.013.

[103] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. S. Pickard, "SUSPECTS: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, no. 6, pp. 773–774, Jan. 2006. DOI: 10.1093/bioinformatics/btk031.

[104] D. Guala, E. Sjölund, and E. L. L. Sonnhammer, "MaxLink: network-based prioritization of genes tightly linked to a disease seed set," *Bioinformatics*, vol. 30, no. 18, pp. 2689–2690, May 2014. DOI: 10.1093/bioinformatics/btu344.

[105] C. Kimmel and S. Visweswaran, "An Algorithm for Network-Based Gene Prioritization That Encodes Knowledge Both in Nodes and in Links," *PLOS ONE*, vol. 8, no. 11, pp. 1–10, Nov. 2013. DOI: 10.1371/journal.pone.0079564. [Online]. Available: https://doi.org/10.1371/journal.pone.0079564.

[106] T. Kacprowski, N. T. Doncheva, and M. Albrecht, "NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules," *Bioinformatics*, vol. 29, no. 11, pp. 1471–1473, Apr. 2013. DOI: 10.1093/bioinformatics/btt164.

[107] T. Yin, S. Chen, X. Wu, and W. Tian, "GenePANDA—a novel network-based gene prioritizing tool for complex diseases," *Scientific Reports*, vol. 7, no. 1, p. 43 258, Mar. 2017. DOI: 10.1038/srep43258.

[108] L.-C. Tranchevent, A. Ardeshirdavani, S. ElShal, D. Alcaide, J. Aerts, D. Auboeuf, and Y. Moreau, "Candidate gene prioritization with Endeavour," *Nucleic Acids Research*, vol. 44, no. W1, W117–W121, Apr. 2016. DOI: 10.1093/nar/gkw365.

[109] M. Franz, H. Rodriguez, C. Lopes, K. Zuberi, J. Montojo, G. D. Bader, and Q. Morris, "GeneMANIA update 2018," *Nucleic Acids Research*, vol. 46, no. W1, W60–W64, Jun. 2018. DOI: 10.1093/nar/gky311.

[110] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, *et al.*, "CDD: NCBI's conserved domain database," *Nucleic Acids Research*, vol. 43, no. D1, pp. D222–D226, Nov. 2014. DOI: 10.1093/nar/gku1221.

[111] J. Mistry, S. Chuguransky, L. Williams, *et al.*, "Pfam: The protein families database in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D412–D419, Oct. 2020. DOI: 10.1093/nar/gkaa913.

[112] M. Blum, H.-Y. Chang, S. Chuguransky, *et al.*, "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Research*, vol. 49, no. D1, pp. D344–D354, Nov. 2020, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa977.

[113] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: the Pathway Interaction Database," *Nucleic acids research*, vol. 37, no. Database issue, pp. D674–D679, Jan. 2009. DOI: 10.1093/nar/gkn653.

[114] P. D. Thomas, M. J. Campbell, A. Kejariwal, *et al.*, "PANTHER: A Library of Protein Families and Subfamilies Indexed by Function," *Genome Research*, vol. 13, no. 9, pp. 2129–2141, 2003. DOI: 10.1101/gr.772403.

[115] I. Rodchenkov, O. Babur, A. Luna, *et al.*, "Pathway Commons 2019 Update: integration, analysis and exploration of pathway data," *Nucleic Acids Research*, vol. 48, no. D1, pp. D489–D497, Oct. 2019. DOI: 10.1093/nar/gkz946.

[116] C. Ogris, D. Guala, M. Kaduk, and E. L. L. Sonnhammer, "FunCoup 4: new species, data, and visualization," *Nucleic Acids Research*, vol. 46, no. D1, pp. D601–D607, Nov. 2017. DOI: 10.1093/nar/gkx1138.

[117] M. Uhlen, P. Oksvold, L. Fagerberg, *et al.*, "Towards a knowledge-based Human Protein Atlas," *Nature Biotechnology*, vol. 28, no. 12, pp. 1248–1250, Dec. 2010. DOI: 10.1038/nbt1210-1248.

[118] M. Uhlen, C. Zhang, S. Lee, *et al.*, "A pathology atlas of the human cancer transcriptome," *Science*, vol. 357, no. 6352, 2017. DOI: 10.1126/science.aan2507.

[119] I. Xenarios, L. Salwínski, X. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, pp. 303–5, Jan. 2002. DOI: 10.1093/nar/30.1.303.

[120] D. L. Wheeler, T. Barrett, D. A. Benson, *et al.*, "Database resources of the National Center for Biotechnology Information," *Nucleic acids research*, vol. 35, no. Database issue, pp. D5–D12, Jan. 2007. DOI: 10.1093/nar/gkl1031.

[121] Q. Xiong, Y. Qiu, and W. Gu, "PGMapper: a web-based tool linking phenotype to genes," *Bioinformatics*, vol. 24, no. 7, pp. 1011–1013, Jan. 2008, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn002.

[122] W. Yu, A. Wulf, T. Liu, M. J. Khoury, and M. Gwinn, "Gene prospector: An evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases," *BMC Bioinformatics*, vol. 9, no. 1, p. 528, Dec. 2008. DOI: 10.1186/1471-2105-9-528.

[123] J.-F. Fontaine, F. Priller, A. Barbosa-Silva, and M. A. Andrade-Navarro, "Génie: literature-based gene prioritization at multi genomic scale," *Nucleic Acids Research*, vol. 39, no. suppl_2, W455–W461, May 2011. DOI: 10.1093/nar/gkr246.

[124] J. Jourquin, D. Duncan, Z. Shi, and B. Zhang, "GLAD4U: deriving and prioritizing gene lists from PubMed literature," *BMC Genomics*, vol. 13, no. 8, S20, Nov. 2012. DOI: 10.1186/1471-2164-13-S8-S20.

[125] Y. Liu, Y. Liang, and D. Wishart, "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more," *Nucleic acids research*, vol. 43, no. W1, W535–W542, Jul. 2015. DOI: 10.1093/nar/gkv383.

[126] S. ElShal, L. Tranchevent, A. Sifrim, A. Ardeshirdavani, J. Davis, and Y. Moreau, "Beegle: From literature mining to disease-gene discovery," *Nucleic acids research*, vol. 44, no. 2, e18–e18, Jan. 2016. DOI: 10.1093/nar/gkv905.

[127] A. Rao, T. Joseph, V. G. Saipradeep, S. Kotte, N. Sivadasan, and R. Srinivasan, "PRIORI-T: A tool for rare disease gene prioritization using MEDLINE," *PLOS ONE*, vol. 15, no. 4, pp. 1–12, Apr. 2020. DOI: 10.1371/journal.pone.0231728.

[128] A. D. Yates, P. Achuthan, W. Akanni, *et al.*, "Ensembl 2020," *Nucleic Acids Research*, vol. 48, no. D1, pp. D682–D688, 2019. DOI: 10.1093/nar/gkz966.

[129] Y. Moreau and L. Tranchevent, "Computational tools for prioritizing candidate genes: Boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 523–536, Aug. 2012. DOI: 10.1038/nrg3253.

[130] K. Moutselos, I. Maglogiannis, and A. Chatziioannou, "GOrevenge: A Novel Generic Reverse Engineering Method for the Identification of Critical Molecular Players, Through the Use of Ontologies," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12, pp. 3522–3527, 2011. DOI: 10.1109/TBME.2011.2164794.

[131] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970. DOI: 10.1016/0022-2836(70)90057-4.

[132] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981. DOI: 10.1016/0022-2836(81)90087-5.

[133] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990. DOI: 10.1016/S0022-2836(05)80360-2.

[134] L. Holm and J. Park, "DaliLite workbench for protein structure comparison," *Bioinformatics*, vol. 16, no. 6, pp. 566–567, Jun. 2000. DOI: 10.1093/bioinformatics/16.6.566.

[135] Y. Ye and A. Godzik, "Flexible structure alignment by chaining aligned fragment pairs allowing twists," *Bioinformatics*, vol. 19, pp. ii246–ii255, Sep. 2003. DOI: 10.1093/bioinformatics/btg1086.

[136] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity : A study in the biomedical domain," *Journal of Biomedical Informatics*, vol. 48, pp. 38–53, 2014. DOI: 10.1016/j.jbi.2013.11.006.

[137] F. Couto and A. Lamurias, "Semantic Similarity Definition," in. Jan. 2019, pp. 870–876. DOI: 10.1016/B978-0-12-809633-8.20401-9.

[138] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'95, Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 448–453.

[139] X. Song, L. Li, P. Srimani, P. Yu, and J. Wang, "Measure the Semantic Similarity of GO Terms Using Aggregate Information Content," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 11, Dec. 2014. DOI: 10.1109/TCBB.2013.176.

[140] G. Mazandu and N. Mulder, "A Topology-Based Metric for Measuring Term Similarity in the Gene Ontology," *Advances in bioinformatics*, vol. 2012, p. 975 783, May 2012. DOI: 10.1155/2012/975783.

[141] F. Couto, M. Silva, and P. Coutinho, "Measuring Semantic Similarity between Gene Ontology Terms," *Data & Knowledge Engineering*, vol. 61, pp. 137–152, Apr. 2007. DOI: 10.1016/j.datak.2006.05.003.

[142] F. Couto and M. Silva, "Disjunctive shared information between ontology concepts: Application to Gene Ontology," *Journal of biomedical semantics*, vol. 2, p. 5, Aug. 2011. DOI: 10.1186/2041-1480-2-5.

[143] G. Mazandu and N. Mulder, "Information Content-Based Gene Ontology Semantic Similarity Approaches: Toward a Unified Framework Theory," *BioMed research international*, vol. 2013, p. 292 063, Sep. 2013. DOI: 10.1155/2013/292063.

[144] C. Pesquita, "Semantic Similarity in the Gene Ontology," in *The Gene Ontology Handbook*, C. Dessimoz and N. Škunca, Eds. New York, NY: Springer New York, 2017, pp. 161–173. DOI: 10.1007/978-1-4939-3743-1_12.

[145] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic Similarity in Biomedical Ontologies," *PLOS Computational Biology*, vol. 5, pp. 1–12, Jul. 2009. DOI: 10.1371/journal.pcbi.1000443.

[146] S. Choi, S.-H. Cha, and C. Tappert, "A Survey of Binary Similarity and Distance," *J. Syst. Cybern. Inf.*, vol. 8, Nov. 2009.

[147] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, ser. ICML '98, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304, ISBN: 1558605568.

[148] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology," *BMC Bioinformatics*, vol. 7, no. 1, p. 302, Jun. 2006. DOI: 10.1186/1471-2105-7-302.

[149] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *Proceedings of the 10th Research on Computational Linguistics International Conference*, Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Aug. 1997, pp. 19–33.

[150] G. Pirró and J. Euzenat, "A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness," in *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, ser. ISWC'10, Shanghai, China: Springer-Verlag, 2010, pp. 615–630.

[151] J. Wang, Z. Du, R. Payattakool, P. Yu, and C.-F. Chen, "A New Method to Measure the Semantic Similarity of GO Terms," *Bioinformatics*, vol. 23, pp. 1274–81, Jun. 2007. DOI: 10.1093/bioinformatics/btm087.

[152] G. K. Mazandu, E. R. Chimusa, M. Mbiyavanga, and N. J. Mulder, *An adaptable gene ontology semantic similarity based functional analysis tool*, version 15.1, Sep. 2015.

[153] G. K. Mazandu and N. J. Mulder, "DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content," *BMC Bioinformatics*, vol. 14, no. 1, p. 284, Sep. 2013. DOI: 10.1186/1471-2105-14-284.

[154] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, Feb. 2010. DOI: 10.1093/bioinformatics/btq064.

[155] T. Koutsandreas, I. Binenbaum, E. Pilalis, I. Valavanis, O. Papadodima, and A. Chatziioannou, "Analyzing and Visualizing Genomic Complexity for the Derivation of the Emergent Molecular Networks," *Int. J. Monit. Surveill. Technol. Res.*, vol. 4, no. 2, pp. 30–49, Apr. 2016. DOI: 10.4018/IJMSTR.2016040103.

[156] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nature protocols*, vol. 4, no. 8, pp. 1184–1191, 2009. DOI: 10.1038/nprot.2009.97.

[157] E. R. Gansner and S. C. North, "An open graph visualization system and its applications to software engineering," *SOFTWARE - PRACTICE AND EXPERIENCE*, vol. 30, no. 11, pp. 1203–1233, 2000.

[158] F. Aurenhammer, "Voronoi Diagrams—a Survey of a Fundamental Geometric Data Structure," *ACM Comput. Surv.*, vol. 23, no. 3, pp. 345–405, Sep. 1991. DOI: 10.1145/116873.116880.

[159] T. Koutsandreas, I. Valavanis, E. Pilalis, and A. Chatziioannou, "An Entropy-based Statistical Workflow Provides Noise-Minimizing Biological Annotation for Muscular Aging," in *2014 8th International Conference on Systems Biology (ISB)*, 2014, pp. 156–163. DOI: 10.1109/ISB.2014.6990749.

[160] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[161] T. Koutsandreas, E. Ladoukakis, E. Pilalis, D. Zarafeta, F. N. Kolisis, G. Skretas, and A. A. Chatziioannou, "ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next Generation Sequencing Data," *Frontiers in Genetics*, vol. 10, p. 469, 2019. DOI: 10.3389/fgene.2019.00469.

[162] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature Biotechnology*, vol. 35, no. 9, pp. 833–844, Sep. 2017. DOI: 10.1038/nbt.3935.

[163] S. G. Tringe and E. M. Rubin, "Metagenomics: DNA sequencing of environmental samples," *Nature Reviews Genetics*, vol. 6, no. 11, pp. 805–814, Nov. 2005. DOI: 10.1038/nrg1709.

[164] G. D. Wu and J. D. Lewis, "Analysis of the Human Gut Microbiome and Association With Disease," *Clinical Gastroenterology and Hepatology*, vol. 11, no. 7, pp. 774–777, Jul. 2013. DOI: 10.1016/j.cgh.2013.03.038.

[165] M. J. Blaser, "Harnessing the power of the human microbiome," *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6125–6126, 2010. DOI: 10.1073/pnas.1002112107.

[166] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature reviews. Genetics*, vol. 13, no. 4, pp. 260–270, Mar. 2012. DOI: 10.1038/nrg3182.

[167] M. J. Blaser and D. Kirschner, "The equilibria that allow bacterial persistence in human hosts," *Nature*, vol. 449, no. 7164, pp. 843–849, Oct. 2007. DOI: 10.1038/nature06198.

[168] E. Afgan, D. Baker, B. Batut, *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic Acids Research*, vol. 46, no. W1, W537–W544, May 2018. DOI: 10.1093/nar/gky379.

[169] E. Boutet, D. Lieberherr, M. Tognolli, *et al.*, "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View," in *Plant Bioinformatics: Methods and Protocols*, D. Edwards, Ed. New York, NY: Springer New York, 2016, pp. 23–54, ISBN: 978-1-4939-3167-5. DOI: 10.1007/978-1-4939-3167-5_2.

[170] T. H. M. J. R. S. Consortium, "A catalog of reference genomes from the human microbiome," *Science (New York, N.Y.)*, vol. 328, no. 5981, pp. 994–999, May 2010. DOI: 10.1126/science.1183605.

[171] D. R. Mende, I. Letunic, O. M. Maistrenko, *et al.*, "proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes," *Nucleic Acids Research*, vol. 48, no. D1, pp. D621–D625, Oct. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gkz1002.

[172] D. H. Parks, C. Rinke, M. Chuvochina, *et al.*, "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life," *Nature Microbiology*, vol. 2, no. 11, pp. 1533–1542, Nov. 2017. DOI: 10.1038/s41564-017-0012-7.

[173] C. Francke, R. J. Siezen, and B. Teusink, "Reconstructing the metabolic network of a bacterium from its genome," *Trends in Microbiology*, vol. 13, no. 11, pp. 550–558, Nov. 2005. DOI: 10.1016/j.tim.2005.09.001.

[174] E. Pitkänen, J. Rousu, and E. Ukkonen, "Computational methods for metabolic reconstruction," *Current Opinion in Biotechnology*, vol. 21, no. 1, pp. 70–77, 2010, Analytical Biotechnology, ISSN: 0958-1669. DOI: 10.1016/j.copbio.2010.01.010.

[175] A. Xenos, T. Koutsandreas, and A. Chatziioannou, *Proposing a unified framework of topological factors in order to refine semantic network analysis on biomedical ontologies*, Bio-Ontologies COSI Track ISMB/ECCB 2019, Jul. 2019. [Online]. Available: https://www.iscb.org/cms_addon/conferences/ismbeccb2019/bioontologies.php.

[176] J. L. Sevilla, V. Segura, A. Podhorski, *et al.*, "Correlation between gene expression and GO semantic similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330–338, 2005. DOI: 10.1109/TCBB.2005.50.

[177] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman, "Assessing semantic similarity for the characterization of human regulatory pathways," *Bioinformatics*, vol. 22, no. 8, pp. 967–973, Feb. 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl042.

[178] S. Craw, "Manhattan Distance," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 639–639, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_506.

[179] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, Jun. 1938, ISSN: 0006-3444. DOI: 10.1093/biomet/30.1-2.81. [Online]. Available: https://doi.org/10.1093/biomet/30.1-2.81.

[180] H. A. Kestler, A. Müller, J. M. Kraus, *et al.*, "Vennmaster: Area-proportional euler diagrams for functional go analysis of microarrays," *BMC Bioinformatics*, vol. 9, no. 1, p. 67, Jan. 2008. DOI: 10.1186/1471-2105-9-67.

[181] L. Cardoso Alves, M. D. Berger, T. Koutsandreas, *et al.*, "Non-apoptotic TRAIL function modulates NK cell activity during viral infection," *EMBO reports*, vol. 21, no. 1, e48789, 2020. DOI: https://doi.org/10.15252/embr.201948789.

[182] L. L. Lanier, "Up on the tightrope: Natural killer cell activation and inhibition," *Nature Immunology*, vol. 9, no. 5, pp. 495–502, May 2008. DOI: 10.1038/ni1581.

[183] C. J. Chan, M. J. Smyth, and L. Martinet, "Molecular mechanisms of natural killer cell activation in response to cellular stress," *Cell Death & Differentiation*, vol. 21, no. 1, pp. 5–14, Jan. 2014. DOI: 10.1038/cdd.2013.26.

[184] S. Paul and G. Lal, "The Molecular Mechanism of Natural Killer Cells Function and Its Importance in Cancer Immunotherapy," *Frontiers in immunology*, vol. 8, pp. 1124–1124, Sep. 2017. DOI: 10.3389/fimmu.2017.01124.

[185] C. Sordo-Bahamonde, S. Lorenzo-Herrero, Á. R. Payer, S. Gonzalez, and A. López-Soto, "Mechanisms of Apoptosis Resistance to NK Cell-Mediated Cytotoxicity in Cancer," *International journal of molecular sciences*, vol. 21, no. 10, p. 3726, May 2020. DOI: 10.3390/ijms21103726.

[186] J. E. Belizário, J. M. Neyra, and M. F. Setúbal Destro Rodrigues, "When and how NK cell-induced programmed cell death benefits immunological protection against intracellular pathogen infection," *Innate immunity*, vol. 24, no. 8, pp. 452–465, Nov. 2018. DOI: 10.1177/1753425918800200.

[187] P. Schneider, D. Olson, A. Tardivel, *et al.*, "Identification of a New Murine Tumor Necrosis Factor Receptor Locus That Contains Two Novel Murine Receptors for Tumor Necrosis Factor-related Apoptosis-inducing Ligand (TRAIL)," *The Journal of biological chemistry*, vol. 278, pp. 5444–54, Mar. 2003. DOI: 10.1074/jbc.M210783200.

[188] G. S. Wu, T. F. Burns, Y. Zhan, E. S. Alnemri, and W. S. El-Deiry, "Molecular Cloning and Functional Analysis of the Mouse Homologue of the KILLER/DR5 Tumor Necrosis Factor-related Apoptosis-inducing Ligand (TRAIL) Death Receptor," *Cancer Research*, vol. 59, no. 12, pp. 2770–2775, 1999.

[189] K. Azijli, B. Weyhenmeyer, G. J. Peters, S. de Jong, and F. A. E. Kruyt, "Non-canonical kinase signaling by the death ligand TRAIL in cancer cells: discord in the death receptor family," *Cell death and differentiation*, vol. 20, no. 7, pp. 858–868, Jul. 2013. DOI: 10.1038/cdd.2013.28.

[190] D. Mérino, N. Lalaoui, A. Morizot, E. Solary, and O. Micheau, "TRAIL in cancer therapy: Present and future challenges," *Expert opinion on therapeutic targets*, vol. 11, pp. 1299–314, Nov. 2007. DOI: 10.1517/14728222.11.10.1299.

[191] H. Ehrhardt, S. Fulda, I. Schmid, J. Hiscott, K.-M. Debatin, and I. Jeremias, "TRAIL induced survival and proliferation in cancer cells resistant towards TRAIL-induced apoptosis mediated by NF-κB," *Oncogene*, vol. 22, no. 25, pp. 3842–3852, Jun. 2003. DOI: 10.1038/sj.onc.1206520.

[192] N. Ishimura, H. Isomoto, S. F. Bronk, and G. J. Gores, "Trail induces cell migration and invasion in apoptosis-resistant cholangiocarcinoma cells," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 290, no. 1, G129–G136, 2006. DOI: 10.1152/ajpgi.00242.2005.

[193] A. Trauzold, D. Siegmund, B. Schniewind, *et al.*, "TRAIL promotes metastasis of human pancreatic ductal adenocarcinoma," *Oncogene*, vol. 25, pp. 7434–9, Dec. 2006. DOI: 10.1038/sj.onc.1209719.

[194] M. Ehrenschwender, D. Siegmund, A. Wicovsky, *et al.*, "Mutant PIK3CA licenses TRAIL and CD95L to induce non-apoptotic caspase-8-mediated ROCK activation," *Cell Death & Differentiation*, vol. 17, no. 9, pp. 1435–1447, Sep. 2010. DOI: 10.1038/cdd.2010.36.

[195] C. Röder, A. Trauzold, and H. Kalthoff, "Impact of death receptor signaling on the malignancy of pancreatic ductal adenocarcinoma," *European Journal of Cell Biology*, vol. 90, no. 6, pp. 450–455, 2011. DOI: https://doi.org/10.1016/j.ejcb.2010.10.008.

[196] S. von Karstedt, A. Conti, M. Nobis, and et al., "Cancer Cell-Autonomous TRAIL-R Signaling Promotes KRAS-Driven Cancer Progression, Invasion, and Metastasis," *Cancer Cell*, vol. 27, no. 4, pp. 561–573, Apr. 2015. DOI: 10.1016/j.ccell.2015.02.014.

[197] *Seven bridges platform*, 2020. [Online]. Available: https://www.sevenbridges.com/.

[198] S. Andrews, *FASTQC. A quality control tool for high throughput sequence data*, 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[199] A. Dobin, C. A. Davis, F. Schlesinger, *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013. DOI: 10.1093/bioinformatics/bts635.

[200] S. Anders, P. T. Pyl, and W. Huber, "HTSeq–a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015. DOI: 10.1093/bioinformatics/btu638.

[201] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, Dec. 2014. DOI: 10.1186/s13059-014-0550-8.

[202] T. Koutsandreas, E. Chevet, A. Chatziioannou, and B. Felden, "Protein homeostasis imprinting across evolution," *bioRxiv*, 2020. DOI: 10.1101/2020.07.19.210591.

[203] W. E. Balch, R. I. Morimoto, A. Dillin, and J. W. Kelly, "Adapting Proteostasis for Disease Intervention," vol. 319, no. 5865, pp. 916–919, 2008. DOI: 10.1126/science.1141448.

[204] "Modeling general proteostasis: Proteome balance in health and disease," *Current Opinion in Cell Biology*, vol. 23, no. 2, pp. 126–134, 2011, Cell regulation. DOI: 10.1016/j.ceb.2010.11.001.

[205] E. Gur, C. Katz, and E. Z. Ron, "All three J-domain proteins of the Escherichia coli DnaK chaperone machinery are DNA binding proteins," *FEBS Letters*, vol. 579, no. 9, pp. 1935–1939, 2005, ISSN: 0014-5793. DOI: https://doi.org/10.1016/j.febslet.2005.01.084.

[206] H. H. Kampinga, J. Hageman, M. J. Vos, *et al.*, "Guidelines for the nomenclature of the human heat shock proteins," *Cell stress & chaperones*, vol. 14, no. 1, pp. 105–111, Jan. 2009. DOI: 10.1007/s12192-008-0068-7.

[207] E. T. Powers, D. L. Powers, and L. M. Gierasch, "FoldEco: A Model for Proteostasis in E.coli," *Cell Reports*, vol. 1, no. 3, pp. 265–276, 2012, ISSN: 2211-1247. DOI: 10.1016/j.celrep.2012.02.011.

[208] R. L. Wiseman, E. T. Powers, J. N. Buxbaum, J. W. Kelly, and W. E. Balch, "An Adaptable Standard for Protein Export from the Endoplasmic Reticulum," *Cell*, vol. 131, no. 4, pp. 809–821, Nov. 2007. DOI: 10.1016/j.cell.2007.10.025.

[209] T. U. Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Nov. 2018. DOI: 10.1093/nar/gky1049.

[210] C. Amid, B. T. F. Alako, V. Balavenkataraman Kadhirvelu, *et al.*, "The European Nucleotide Archive in 2019," *Nucleic Acids Research*, vol. 48, no. D1, pp. D70–D76, Nov. 2019. DOI: 10.1093/nar/gkz1063.

[211] P. Gaudet and C. Dessimoz, "Gene Ontology: Pitfalls, Biases, and Remedies," in *The Gene Ontology Handbook*, C. Dessimoz and N. Škunca, Eds. New York, NY: Springer New York, 2017, pp. 189–205. DOI: 10.1007/978-1-4939-3743-1_14.

[212] B. Lobb, B. J.-M. Tremblay, G. Moreno-Hagelsieb, and A. C. Doxey, "An assessment of genome annotation coverage across the bacterial tree of life," *Microbial genomics*, vol. 6, no. 3, e000341, Mar. 2020. DOI: 10.1099/mgen.0.000341.

[213] J. H. and W. Jr., "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. DOI: 10.1080/01621459.1963.10500845.

[214] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: The primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, Nov. 1977. DOI: 10.1073/pnas.74.11.5088.

[215] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, Jun. 1990. DOI: 10.1073/pnas.87.12.4576.

[216] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Science*, vol. 27, no. 1, pp. 135–145, 2018. DOI: 10.1002/pro.3290.

[217] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, May 2006. DOI: 10.1093/bioinformatics/btl158.

[218] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, W29–W37, Jul. 2011. DOI: 10.1093/nar/gkr367.

[219] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 410–420.

[220] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

[221] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Aug. 2005. DOI: 10.1007/978-1-4757-2711-1.

[222] D. E. Gordon, G. M. Jang, M. Bouhaddou, *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," *Nature*, vol. 583, no. 7816, pp. 459–468, Jul. 2020. DOI: 10.1038/s41586-020-2286-9.

[223] D. Sicari, A. Chatziioannou, T. Koutsandreas, R. Sitia, and E. Chevet, "Role of the early secretory pathway in SARS-CoV-2 infection," *The Journal of cell biology*, vol. 219, no. 9, e202006005, Sep. 2020. DOI: 10.1083/jcb.202006005.

[224] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet. Infectious diseases*, vol. 20, no. 5, pp. 533–534, May 2020. DOI: 10.1016/S1473-3099(20)30120-1.

[225] M. Roser, H. Ritchie, E. Ortiz-Ospina, and J. Hasell, "Coronavirus Pandemic (COVID-19)," *Our World in Data*, 2020, https://ourworldindata.org/coronavirus.

[226] D. Wang, B. Hu, C. Hu, *et al.*, "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China," *JAMA*, vol. 323, no. 11, pp. 1061–1069, Mar. 2020. DOI: 10.1001/jama.2020.1585.

[227] S. Su, G. Wong, W. Shi, *et al.*, "Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses," *Trends in Microbiology*, vol. 24, no. 6, pp. 490–502, 2016, ISSN: 0966-842X. DOI: https://doi.org/10.1016/j.tim.2016.03.003.

[228] R. Lu, X. Zhao, J. Li, and et al., "Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding," *Lancet (London, England)*, vol. 395, no. 10224, pp. 565–574, Feb. 2020. DOI: 10.1016/S0140-6736(20)30251-8.

[229] A. R. Fehr and S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis," *Methods in molecular biology (Clifton, N.J.)*, vol. 1282, pp. 1–23, 2015. DOI: 10.1007/978-1-4939-2438-7_1.

[230] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," *Journal of medical virology*, vol. 92, no. 4, pp. 418–423, Apr. 2020. DOI: 10.1002/jmv.25681.

[231] Y. Huang, C. Yang, X.-f. Xu, W. Xu, and S.-w. Liu, "Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19," *Acta Pharmacologica Sinica*, vol. 41, no. 9, pp. 1141–1149, Sep. 2020. DOI: 10.1038/s41401-020-0485-4.

[232] J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, and F. Li, "Cell entry mechanisms of SARS-CoV-2," *Proceedings of the National Academy of Sciences*, vol. 117, no. 21, pp. 11727–11734, 2020. DOI: 10.1073/pnas.2003138117.

[233] M. Hoffmann, H. Kleine-Weber, S. Schroeder, *et al.*, "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell*, vol. 181, no. 2, 271–280.e8, Apr. 2020. DOI: 10.1016/j.cell.2020.02.052.

[234] I.-C. Huang, B. J. Bosch, F. Li, *et al.*, "SARS Coronavirus, but Not Human Coronavirus NL63, Utilizes Cathepsin L to Infect ACE2-expressing Cells," *Journal of Biological Chemistry*, vol. 281, no. 6, pp. 3198–3203, Feb. 2006. DOI: 10.1074/jbc.M508381200.

[235] B. Coutard, C. Valle, X. de Lamballerie, B. Canard, N. G. Seidah, and E. Decroly, "The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade," *Antiviral research*, vol. 176, pp. 104742–104742, Apr. 2020. DOI: 10.1016/j.antiviral.2020.104742.

[236] V. Mody, J. Ho, S. Wills, *et al.*, "Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents," *Communications Biology*, vol. 4, no. 1, p. 93, Jan. 2021. DOI: 10.1038/s42003-020-01577-x.

[237] Y. Gao, L. Yan, Y. Huang, *et al.*, "Structure of the RNA-dependent RNA polymerase from COVID-19 virus," *Science*, vol. 368, no. 6492, pp. 779–782, 2020. DOI: 10.1126/science.abb7498.

[238] T. S. Fung and D. X. Liu, "Coronavirus infection, ER stress, apoptosis and innate immunity," *Frontiers in Microbiology*, vol. 5, p. 296, 2014. DOI: 10.3389/fmicb.2014.00296.

[239] C. Wu, Y. Liu, Y. Yang, *et al.*, "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods," *Acta Pharmaceutica Sinica B*, vol. 10, Feb. 2020. DOI: 10.1016/j.apsb.2020.02.008.

[240] B. Teng, C. Zhao, X. Liu, and Z. He, "Network inference from AP-MS data: computational challenges and solutions," *Briefings in Bioinformatics*, vol. 16, no. 4, pp. 658–674, Nov. 2014. DOI: 10.1093/bib/bbu038. [Online]. Available: https://doi.org/10.1093/bib/bbu038.

[241] S. Jäger, P. Cimermancic, N. Gulbahce, *et al.*, "Global landscape of HIV-human protein complexes," *Nature*, vol. 481, no. 7381, pp. 365–370, Dec. 2011, ISSN: 1476-4687. DOI: 10.1038/nature10719.

[242] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper, "Defining the human deubiquitinating enzyme interaction landscape," *Cell*, vol. 138, no. 2, pp. 389–403, Jul. 2009, ISSN: 1097-4172. DOI: 10.1016/j.cell.2009.04.042.

[243] G. Teo, G. Liu, J. Zhang, A. Nesvizhskii, A.-C. Gingras, and H. Choi, "SAINTexpress: Improvements and additional features in Significance Analysis of INTeractome software," *Journal of proteomics*, vol. 100, Oct. 2013. DOI: 10.1016/j.jprot.2013.10.023.

[244] H. R. Ramage, G. R. Kumar, E. Verschueren, *et al.*, "A Combined Proteomics/Genomics Approach Links Hepatitis C Virus Infection with Nonsense-Mediated mRNA Decay," *Molecular Cell*, vol. 57, no. 2, pp. 329–340, 2015, ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2014.12.028.

[245] Z. H. Davis, E. Verschueren, G. M. Jang, *et al.*, "Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes," *Molecular cell*, vol. 57, no. 2, pp. 349–360, Jan. 2015, ISSN: 1097-4164. DOI: 10.1016/j.molcel.2014.11.026.

[246] M. Eckhardt, W. Zhang, A. M. Gross, *et al.*, "Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network," *Cancer discovery*, vol. 8, no. 11, pp. 1474–1489, Nov. 2018, ISSN: 2159-8290. DOI: 10.1158/2159-8290.CD-17-1018.

[247] J. Batra, J. F. Hultquist, D. Liu, *et al.*, "Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication," *Cell*, vol. 175, no. 7, 1917–1930.e13, Dec. 2018. DOI: 10.1016/j.cell.2018.08.044.

[248] P. S. Shah, N. Link, G. M. Jang, *et al.*, "Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis," *Cell*, vol. 175, no. 7, 1931–1945.e18, Dec. 2018. DOI: 10.1016/j.cell.2018.11.028.

[249] M. Li, J. R. Johnson, B. Truong, *et al.*, "Identification of antiviral roles for the exon-junction complex and nonsense-mediated decay in flaviviral infection," *Nature microbiology*, vol. 4, no. 6, pp. 985–995, Jun. 2019, ISSN: 2058-5276. DOI: 10.1038/s41564-019-0375-z.

[250] J. Diep, Y. S. Ooi, A. W. Wilkinson, *et al.*, "Enterovirus pathogenesis requires the host methyltransferase SETD3," *Nature microbiology*, vol. 4, no. 12, pp. 2523–2537, Dec. 2019, ISSN: 2058-5276. DOI: 10.1038/s41564-019-0551-1.

[251] Y.-Y. Zheng, Y.-T. Ma, J.-Y. Zhang, and X. Xie, "COVID-19 and the cardiovascular system," *Nature Reviews Cardiology*, vol. 17, no. 5, pp. 259–260, May 2020. DOI: 10.1038/s41569-020-0360-5.

[252] K. J. Clerkin, J. A. Fried, J. Raikhelkar, *et al.*, "Covid-19 and cardiovascular disease," *Circulation*, vol. 141, no. 20, pp. 1648–1655, 2020. DOI: 10.1161/CIRCULATIONAHA.120.046941.

[253] L. Ferini-Strambi and M. Salsone, "COVID-19 and neurological disorders: are neurodegenerative or neuroimmunological diseases more vulnerable?" *Journal of neurology*, vol. 268, no. 2, pp. 409–419, Feb. 2021. DOI: 10.1007/s00415-020-10070-8.

[254] A. Verkhratsky, Q. Li, S. Melino, G. Melino, and Y. Shi, "Can COVID-19 pandemic boost the epidemic of neurodegenerative diseases?" *Biology Direct*, vol. 15, no. 1, p. 28, Nov. 2020. DOI: 10.1186/s13062-020-00282-3.

[255] M. Taquet, J. R. Geddes, M. Husain, S. Luciano, and P. J. Harrison, "6-month neurological and psychiatric outcomes in 236379 survivors of COVID-19: a retrospective cohort study using electronic health records," *The Lancet Psychiatry*, Apr. 2021. DOI: 10.1016/S2215-0366(21)00084-5.

[256] G. Li and M. C. Marlin, "Rab family of GTPases," *Methods in molecular biology (Clifton, N.J.)*, vol. 1298, pp. 1–15, 2015, ISSN: 1940-6029. DOI: 10.1007/978-1-4939-2569-8_1.

[257] P. Spearman, "Viral interactions with host cell Rab GTPases," *Small GTPases*, vol. 9, no. 1-2, pp. 192–201, Mar. 2018, ISSN: 2154-1256. DOI: 10.1080/21541248.2017.1346552.

[258] V. Nofrini, D. Di Giacomo, and C. Mecucci, "Nucleoporin genes in human diseases," *European Journal of Human Genetics*, vol. 24, no. 10, pp. 1388–1395, Oct. 2016, ISSN: 1476-5438. DOI: 10.1038/ejhg.2016.25.

[259] K. Kato, D. K. Ikliptikawati, A. Kobayashi, H. Kondo, K. Lim, M. Hazawa, and R. W. Wong, "Overexpression of SARS-CoV-2 protein ORF6 dislocates RAE1 and NUP98 from the nuclear pore complex," *Biochemical and biophysical research communications*, vol. 536, pp. 59–66, Jan. 2021. DOI: 10.1016/j.bbrc.2020.11.115.

[260] L. Miorin, T. Kehrer, M. T. Sanchez-Aparicio, *et al.*, "SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling," *Proceedings of the National Academy of Sciences*, vol. 117, no. 45, pp. 28 344–28 354, 2020, ISSN: 0027-8424. DOI: 10.1073/pnas.2016650117.

[261] C. J. Sigrist, A. Bridge, and P. Le Mercier, "A potential role for integrins in host cell entry by SARS-CoV-2," *Antiviral research*, vol. 177, pp. 104 759–104 759, May 2020. DOI: 10.1016/j.antiviral.2020.104759.

[262] S. Yan, H. Sun, X. Bu, and G. Wan, "New Strategy for COVID-19: An Evolutionary Role for RGD Motif in SARS-CoV-2 and Potential Inhibitors for Virus Infection," *Frontiers in Pharmacology*, vol. 11, p. 912, 2020. DOI: 10.3389/fphar.2020.00912.

[263] M. Santerre, S. P. Arjona, C. N. Allen, N. Shcherbik, and B. E. Sawaya, "Why do SARS-CoV-2 NSPs rush to the ER?" *Journal of neurology*, pp. 1–10, Sep. 2020. DOI: 10.1007/s00415-020-10197-8.

[264] V. P. Sebastián, G. A. Salazar, I. Coronado-Arrázola, *et al.*, "Heme Oxygenase-1 as a Modulator of Intestinal Inflammation Development and Progression," *Frontiers in Immunology*, vol. 9, p. 1956, 2018. DOI: 10.3389/fimmu.2018.01956.

[265] M. Dattilo, "The role of host defences in Covid 19 and treatments thereof," *Molecular Medicine*, vol. 26, no. 1, p. 90, Sep. 2020. DOI: 10.1186/s10020-020-00216-9.

[266] N. Batra, C. De Souza, J. Batra, A. G. Raetz, and A.-M. Yu, "The HMOX1 Pathway as a Promising Target for the Treatment and Prevention of SARS-CoV-2 of 2019 (COVID-19)," *International journal of molecular sciences*, vol. 21, no. 17, p. 6412, Sep. 2020. DOI: 10.3390/ijms21176412.

[267] J. A. Keating and R. Striker, "Phosphorylation events during viral infections provide potential therapeutic targets," *Reviews in medical virology*, vol. 22, no. 3, pp. 166–181, May 2012. DOI: 10.1002/rmv.722.

[268] E. Weisberg, A. Parent, P. L. Yang, *et al.*, "Repurposing of Kinase Inhibitors for Treatment of COVID-19," *Pharmaceutical research*, vol. 37, no. 9, pp. 167–167, Aug. 2020. DOI: 10.1007/s11095-020-02851-7.

[269] G. Garcia, A. Sharma, A. Ramaiah, *et al.*, "Antiviral Drug Screen of Kinase inhibitors Identifies Cellular Signaling Pathways Critical for SARS-CoV-2 Replication," *bioRxiv*, 2020. DOI: 10.1101/2020.06.24.150326.

[270] A. Subramanian, R. Narayan, S. M. Corsello, *et al.*, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell*, vol. 171, no. 6, 1437–1452.e17, Nov. 2017. DOI: 10.1016/j.cell.2017.10.049.

[271] S. Yan, R.-H. Liu, H.-Z. Jin, X.-R. Liu, J. Ye, L. Shan, and W.-D. Zhang, ""Omics" in pharmaceutical research: overview, applications, challenges, and future perspectives," *Chinese Journal of Natural Medicines*, vol. 13, no. 1, pp. 3–21, 2015, ISSN: 1875-5364. DOI: https://doi.org/10.1016/S1875-5364(15)60002-4.

[272] B. Amer and E. E. K. Baidoo, "Omics-Driven Biotechnology for Industrial Applications," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 30, 2021. DOI: 10.3389/fbioe.2021.613307.