

Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

**ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ**

**<A Brief Review OF Data Analytics in Bioinformatics: A Machine and Deep Learning Perspective>**

<Ιωάννης-Ορέστης Παγδάτογλου>

Α.Μ. <1067433>

<Βλάσιος-Παναγιώτης Παναγιώτου>

Α.Μ. <1067517>

Πάτρα, <5ο>

Πάτρα, <4ο>

\*\* θα θέλαμε να εξεταστούμε στις 3 Ιουλίου 2023 \*\*

**Εισαγωγή**

Τα Bioinformatics εισάγουν διάφορα ζητήματα βιολογίας στην επιστήμη των υπολογιστών. Αναλύουν βιολογικά δεδομένα όπως πρωτεΐνες, κύτταρα, γονίδια, και ασχολούνται με την υπολογιστική τους αξιοποίηση πάνω σε οικολογικά περιβάλλοντα, γενετική μηχανική, δημιουργία φαρμάκων κ.ά. Δηλαδή μπορούμε να τα χρησιμοποίησουμε για να κάνουμε προβλέψεις, μοντελοποιήσεις, οπτικοποιήσεις ή και για σχεδίαση.

Ουσιαστικά αποτελούν μια τεχνική που μας βοηθάει να φτίαξουμε μαθηματικά μοντέλα με την αξιοποιήση των υπολογιστών για να επεξεργαστούμε διάφορα βιολογικά δεδομένα. Συνήθως αύτου του είδους τα δεδομένα έχουν μεγάλο μέγεθος και συσχετίζονται με περίπλοκους τρόπους καθηστώντας την χρήση Bioinformatics αναγκαία για να δημιουργηθούν εργαλεία με τα οποία θα μπορούμε να τα ερμηνεύσουμε.

Υπάρχει μεγάλη επιρροή της βιοπληροφορικής σε πολλές περιοχές της επιστήμης της βιολογίας:

1. **Γενετική**

γονιδιακή πρόβλεψη, κατανόηση της γονιδιακής ροής,

γονιδιακή αλληλουχία, ερμηνεία γονιδιωμάτων και παρατηρούμενων μεταλλάξεων τους.

1. **Επεξεργασία σήματος και εικόνας**

πειραματική μοριακή βιολογία για την εξόρυξη χρήσιμων πληροφοριών από μεγάλα ποσότητα ακατέργαστων δεδομένων

1. **Εξόρυξη κειμένου**

βοηθούν στη μελέτη και οργανώνουν βιολογικά δεδομένα πολλών βιολογικών βιβλιογραφιών, που οδηγούν σε επέκταση πολλών βιολογικών και γονιδιακών οντολογιών.

1. **Μελέτη πρωτεινών και γονιδιακής έκφρασης**
2. **Βιολογία συστημάτων**

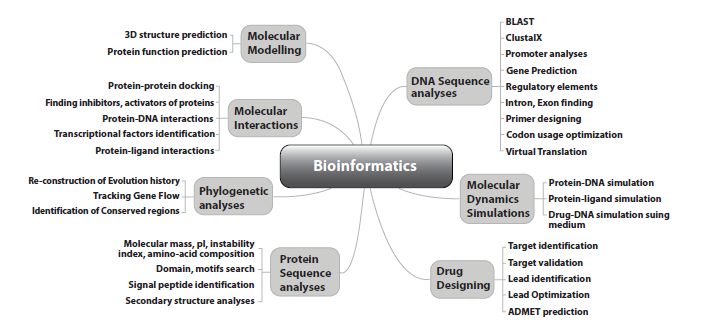
αναλύουν και καταγράφουν τη βιολογικά διαδικασία και τα βιολογικά δίκτυα. Βοηθά επίσης στην απόκτηση γνώσεων σχετικά με το εξελικτικά χαρακτηριστικά της μοριακής βιολογίας για ανάλυση, σύγκριση και ερμηνεία γενετικών και γονιδιωματικών δεδομένων

1. **Πρόβλεψη τρισδιάστατης δομής διαφόρων μακρομορίων όπως RNA, DNA , πρωτείνες.**
2. **Δομική βιολογία**

Ανάλυση και προσομοίωση διαφορετικών βιο-μοριακών αλληλεπιδράσεων

**Τομείς εφαρμογών των Bioinformatics**

Παρακάτω ακολουθεί ένα σχήμα που συνοψίζει τα παραπάνω, και περιγράφει διάφορες εφαρμογές της βιοπληροφορικής σε κάθε τομέα.



**Εφαρμογές Bioinformatics στην πραγματική ζωή**

Υγεία

• Ανάλυση του γονιδιώματος για την ανίχνευση και ταξινόμηση της ασθένειας και για τη βελτίωση της ιατρικής διάγνωσης.

• Μελέτη γονιδίων ή πρωτεϊνών που σχετίζονται με τον παθογόνο παράγοντα και την παθογένεια μιας ασθένειας*.*

• Υπολογισμός της πιθανότητας θανάτου ενός ασθενούς χρησιμοποιώντας κλινικά και γενετικά δεδομένα.

Ανάπτυξη φαρμάκων

• Δημιουργία και γρήγορος έλεγχος φαρμάκων βρίσκοντας την συσχέτιση μεταξύ των γονιδίων και των αντίστοιχω ασθενειών.

• Ανάλυση, πρόβλεψη και ερμηνεία κλινικών και προκλινικών δεδομένων.

• Στόχευση της πηγής της ασθένειας, όχι αντιμετώπιση των συμπτωμάτων.

Μοριακά ιατρική

• Υπολογισμός της ευαλωτότητας ενός ανθρώπου σε μια ασθένεια με μελέτη των γενετικών μηχανισμών για τον εντοπισμό συγκεκιμένων γονιδίων.

• Εύρεση και διόρθωση μοριακών και γονιδιακών σφαλμάτων σχετιζόμενα με ασθένειες.

Προληπτική ιατρική

• Εντοπισμός της πηγής μιας ασθένειας.

• Ανάπτυξη προληπτικών φαρμάκων.

Γονιδιακή θεραπεία

• Τροποποίηση της έκφρασης γονιδίων ενός ατόμου με σκοπό τη θεραπεία, την αντιμετώπιση και την πρόληψη ασθενειών.

Καθαρισμός αποβλήτων

• Εντοπισμός βακτηρίων όπως το Deinococcus radiodurans, τα οποία βοηθούν στον καθαρισμό αποβλήτων.

Μελέτη της κλιματικής αλλαγής

• Μελέτη οργανισμών που χρησιμοποιούν το διοξείδιο του άνθρακα ως κύρια πηγή ενέργειας.

Βιοτεχνολογία

• Ταυτοποίηση οργανισμών και μικρο-οργανισμών που μπορούν να είναι χρήσιμοι στην βιομηχανία και στην παραγωγή τροφίμων όπως το *Lactococcus lactis* .

Βελτίωση σοδειάς

• Δημιουργία φυτών που είναι πιο ανθεκτικά στα έντομα και πιο παραγωγικά μέσω της μελέτης της γενετικής τους δομής.

Αντοχή στα έντομα

• Δημιουργία βακτηρίων όπως το *Bacillus thuringiensis* που φτίαχνουν πρωτεΐνες που είναι τοξικές σε ορισμένα έντομα.

Ανάπτυξη ποικιλιών με αντοχή στην ξηρασία

• Ανάπτυξη φυτών με αντοχή στην αλαλικότητα του εδάφους, τοξικότητες του σιδήρου και στην έλειψη του νερού με σκοπό την δημιουργία γεωργικής γης σε μέρη με χαμηλή ποιότητα εδάφους.

Συγκριτικές μελέτες

• Ανάλυση και σύγκριση γενετικής πληροφορίας από διαφορετικά είδη για να μελετηθούν οι βιοχημικές λειτουργίες γονιδίων σε διαφορετικούς οργανισμούς.

**Θέματα/προβλήματα των Bioinformatics**

Οι εφαρμογές που αναλύθηκαν παραπάνω έχουν και αρκετές προκλήσεις και τεχνικές δυσκολίες , οι οποίες σχετίζονται με τα δεδομένα αυτά καθ’ αυτά ή τις συσκευές που χρησιμοποιούνται για τη συλλογή και την ανάλυση των δεδομένων.

Παρακάτω παραθέτουμε ορισμένες προκλήσεις – ζητήματα – θέματα όσον αφορά τη δομή των μακρομορίων και την ανάλυση των αλληλουχιών των μακρομορίων αυτών:

• Θέματα που σχετίζονται με την δομή:

Η μελέτη του DNA και των πρωτεϊνών περιλαμβάνει προβλήματα όπως η πρόβλεψη της δομής τους στον τρισδιάστατο χώρο, η ευθυγράμμιση τους και η ανάλυσή της δομής που καθιστούν δύσκολο πρόβλημα.Ειδικότερα, η πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών από την αλληλουχία τους μπορεί να επιλυθεί με την εφαρμογή Τεχνητών Νευρωνικών Δικτύων (ANN). Τα περισσότερα βιοογικά δίκτυα, όπως τα δίκτυα αλληλεπίδρασης πρωτεϊνών, τα γονιδιακά ρυθμιστικά δίκτυα κ.λπ., είναι δύσκολο να ερμηνευτούν και να δημιουργηθούν λόγω της πολυπλοκότητας του βιολογικού συστήματος. Έτσι, χρησιμοποιώντας μεθόδους της θεωρίας γραφημάτων, αυτά τα δίκτυα αναπαρίστανται σε γραφήματα, κάτι που καθιστά την κατηγοριοποίηση πολύ δύσκολη χρησιμοποιώντας παραδοσιακές μεθόδους.

• Θέματα που σχετίζονται με την ανάλυση αλληλουχιών:

Η κατηγοριοποίηση των αλληλουχιών RNA, πρωτεϊνών και DNA αποτελεί πρόκληση λόγω των διαφορών και ομοιότητας πολλών οργανισμών.

Πέρα από τη συζήτηση για τα βιολογικά δεδομένα και τις τεχνικές δυσκολίες που προκύπτουν για τη συλλογή και τη διαχείριση τους, στην επιστήμη της βιοπληροφορικής υπάρχει και η εξίσου σημαντική παράμετρος της μαθηματικής μοντελοποίησης και της κατασκευής μαθηματικών μοντέλων για την αντιμετώπιση συγκεκριμένων εφαρμογών (με την αξιοποίηση των υπολογιστών για να επεξεργαστούμε διάφορα βιολογικά δεδομένα, όπως αναφέραμε στην εισαγωγή της αναφοράς μας)

Παρακάτω αναφέρουμε κάποιες προϋποθέσεις που είναι αναγκαίες για τη κατασκευή μαθηματικών μοντέλων βιοπληροφορικής:

**Προ-Επεξεργασία Δεδομένων και η Αναγκαιότητά της**

Μετά τη συλλογή των δεδομένων μας τα οποία είναι σε ακατέργαστη μορφή , πραγματοποιούμε μια προ-επεξεργασία ώστε να τα βελτιώσουμε ποιοτικά και να τα αποθορυβοποιήσουμε , το οποίο κατ’ επέκταση μας βοηθάει στο να έχουμε εν τέλει αυξημένη απόδοση για το μοντέλο κατηγοριοποίησης και μειωμένο χρόνο στην μετέπειτα εξόρυξη γνώσης που θέλουμε να εκτελέσουμε.

**Classification - κατηγοριοποίηση βιολογικών δεδομένων**

Όπως αναφέραμε και αμέσως από πάνω στην προ-επεξεργασία των δεδομένων, η κατηγοριοποίηση (classification) παίζει σημαντικό ρόλο για την εξόρυξη γνώσης και πληροφορίας.

Το classification είναι μια διαδικασία με την οποία τα δεδομένα οργανώνονται σε διάφορες κατηγορίες καθορίζοντας μια κλάση για κάθε στοιχείο στη βάση δεδομένων.

Η κατηγοριοποίηση βοηθάει ιδιαίτερα στην αποδοτική και γενικότερα ορθή ανάλυση των δεδομένων μας, αλλά επίσης και στη κατασκευή μοντέλων που ορίζουν τις κλάσεις δεδομένων και δύναται να κάνουν πρόβλεψη για μελλοντικές συμπεριφορές στα δεδομένα μας.

**Machine Learning και Bioinformatics**

Το Machine Learning – ML, όπως και το Deep Learning ως κομμάτι του ML, είναι ένας κλάδος της τεχνητής νοημοσύνης (AI) και της επιστήμης των υπολογιστών που επικεντρώνεται στη χρήση δεδομένων και αλγορίθμων για να προσομοιώσει τον τρόπο με τον οποίο οι άνθρωποι μαθαίνουν, βελτιώνοντας σταδιακά την ακρίβειά του εκάστοτε μοντέλου.

Αυτές οι τεχνικές έχουν την δυνατότητα να ‘μαθαίνουν’ βασιζόμενες σε ένα αρχικό σύνολο δεδομένων (training set) και με βάση αυτό να κάνουν προβλέψεις. Αυτό σημαίνει ότι το παραγώμενο μοντέλο μπορεί μόνο του χωρίς την παρέμβαση του ανθρώπου να εξάγει συμπεράσματα.

Οι ερευνητές σήμερα, και γενικότερα κατά τη διάρκεια των τελευταίων χρόνων, έχουν τη δυνατότητα χρήσης αυτών των τεχνικών χάριν στη πληθώρα διαθέσιμων βιολογικών δεδομένων και την αυξημένη υπολογιστική ισχύ των συστημάτων.

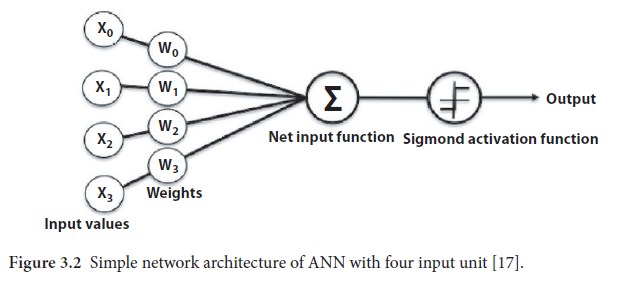
Υπάρχουν εφαρμογές αυτής της τεχνολογίας σε πολλούς βιολογικούς τομείς, όπως οι μικροσυστοιχίες, η συστημική βιολογία, η γενομική, η πρωτεομικροσκοπία, η διάγνωση εγκεφαλικού επεισοδίου και η ανάκτηση πληροφοριών από κείμενο κ.λπ. [8].

Το Machine Learning μπορεί επίσης να κατηγοριοποιηθεί σε επιβλεπόμενη μάθηση (supervised learning), μη επιβλεπόμενη μάθηση (unsupervised learning) και ημι-επιβλεπόμενη μάθηση (semi-supervised learning).

Ωστόσο, υπάρχουν πολλά μοντέλα/τεχνικές μηχανικής μάθησης διαθέσιμα σήμερα. Το αποδοτικότερο για αναγνώριση, επιλογή κατηγοριοποίηση και γονιδιακής πρόβλεψης σε αλληλουχίες DNA , σύμφωνα με το πέιπερ «Data Analytics in Bioinformatics - 2021 - Satpathy - A Critical Review on the Application of Artificial Neural Network in», είναι το Artificial Neural Network

Το Artificial Neural Network (ΑΝΝ) είναι τεχνική μηχανικής μάθησης, Συγκεκριμένα, είναι ένα σύστημα που προσομοιάζει στον ανθρώπινο εγκέφαλο, ο οποίος περιέχει νευρώνες, και δύναται να διαχειριστεί τη πληροφορία με σύνθετους χειρισμούς επί των σύνθετων χαρακτηριστικών που διέπουν/χαρακτηρίζουν τα δεδομένα μας.

Μια ενδεικτική απλή απεικόνισή της δομής του ΑΝΝ (με 1 input layer) είναι η εξής:

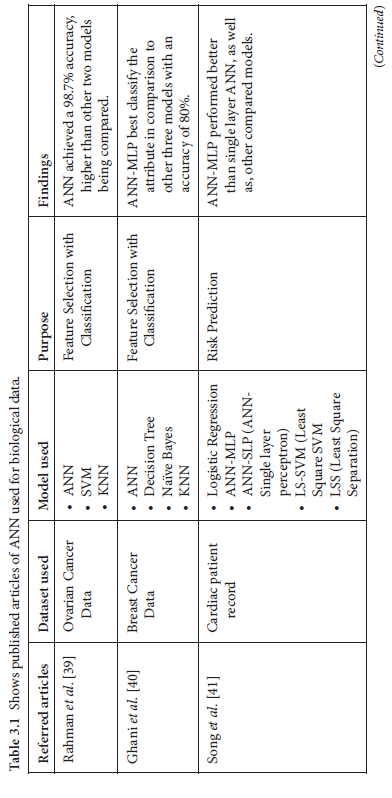


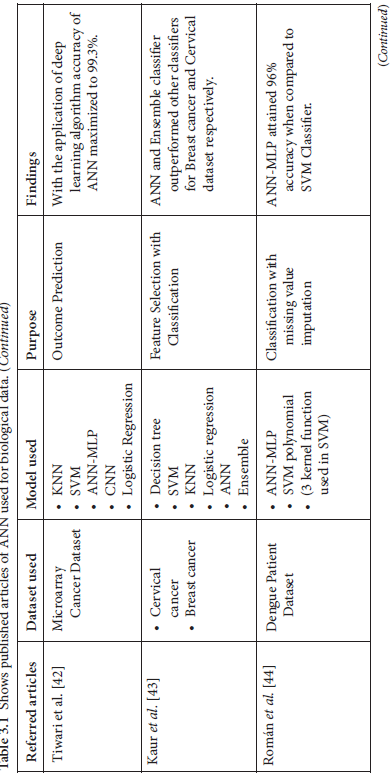
Τα ΑΝΝs χωρίζονται σε διάφορους τύπους ανάλογα με τη δομή τους , με τη πυκνότητα των νευρώνων, τον αριθμό των layers του ANN κλπ. Η πιο συνήθης δομή είναι η τριών στρωμάτων (3-layer) όπως στη φωτό από πάνω, που έχει ένα input layer, ένα ή περισσότερα hidden layers και ένα output layer.

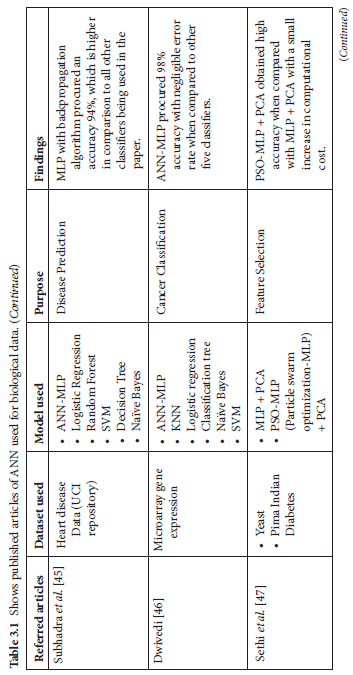
Εφαρμογές των ΑΝΝ στη βιοπληροφορική

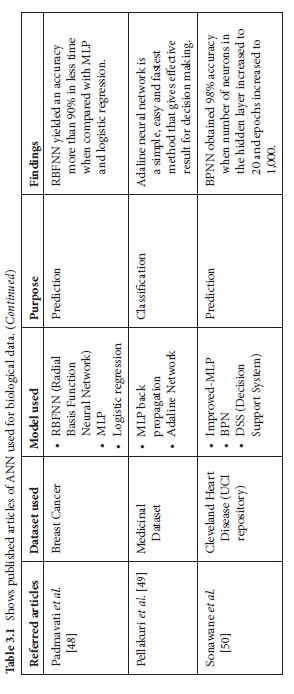
* Ευθυγράμμιση/στοίχιση αλληλουχιών DNA και RNA (DNA-RAN alignment)
* Επεξεργασία σήματος και εικόνας, για πρόβλεψη καρδιοπαθειών, καρκίνου πνεύμονα και ήπατος
* Στα προβλήματα μελέτης και αποκωδικοποίησης γονιδίων
* Αναγνώριση κωδικοποιούμενων και μη-κωδικοποιούμενων περιοχών (εξωνίων και εσωνίων) στις αλληλουχίες DNA
* Αναγνώριση και κατηγοριοποίηση ακολουθιακών δεδομένων
* Πρόβλεψη δευτεροταγούς και τριτοταγούς δομής πρωτεϊνών από διαφορετικές αλληλουχίες αμινοξέων
* Γονιδιακή έκφραση
* Επίβλεψη/παρακολούθηση θεραπείας ασθενών βασισμένη σε αλληλουχίες DNA και πρόβλεψη ποσοστού επιβίωσης ασθενών (πχ ασθενών με νευροβλάστωμα)
* Πρόβλεψη εμφάνισης καρκίνου μαστού από συγκεκριμένο dataset με χρήση ΑΝΝ
* Αναγνώριση γονιδίων που σχετίζονται με εμφάνιση πάθησης στο μυελό των οστών μέσα από αξιοποίηση ενός τεράστιου dataset ασθενών με ΑΝΝ
* Ταξινόμηση γονιδίων που σχετίζονται με καρκίνο ουροδόχου κύστης μέσα από ιστοπαθολογικές εικόνες και απεικονίσεις (MRI, PET-CT Scans κλπ)

Σύγκριση επιδόσεων ανάμεσα σε ΑΝΝs και σε ευρέως γνωστούς αλγορίθμους μηχανικής μάθησης





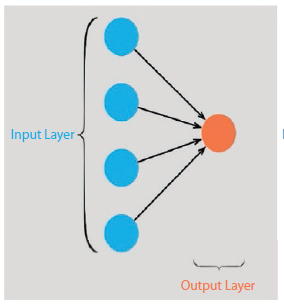




Στη συνέχεια. παραθέτουμε κάποια από τα ΑΝΝ μοντέλα τα οποία χρησιμοποιούνται ευρέως ως μοντέλα επιβλεπόμενης μάθησης (supervised learning).

Μετά από την παράθεση, θα αναλύσουμε τι εστί ενισχυτική μάθηση (supervised learning).

• Perceptron Network (one layer):

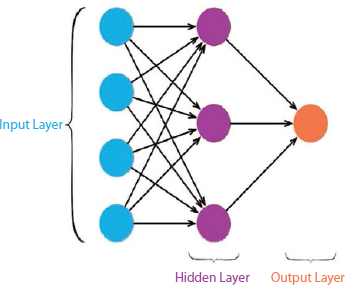


Αποτελεί μοντέλο δυαδικού ταξινομητή που διαχωρίζει τα δεδομένα εισόδου σε 2 κατηγορίες.

Χρησιμοποιείται για γραμμικά διαχωρίσιμα (linearly seperable) προβλήματα

Για μη-γραμμικά διαχωρίσιμα, και γενικότερα για πιο σύνθετα προβλάματα, χρησιμοποιείται η multilayer εκδοχή του, που αναλύουμε αμέσως μετά

• Multilayer Perceptron (MLP) Network:



Χρήση σε multi-class προβλήματα ταξινόμησης και μη-γραμμικά διαχωρίσιμα.

Περιέχει τουλάχιστον ένα hidden layer (ενώ το perceptron network είχε μόνο ένα)

Κάθε κόμβος συνδέεται με κάθε άλλον κόμβο στο επόμενο layer του network

Τι εστί όμως ενισχυτική μάθηση (supervised learning);

Η Βαθεία Μάθηση (Deep Learning) αποτελεί μια τάξη προσεγγίσεων μηχανικής μάθησης («Deep learning, as a class of machine learning approaches,»), δηλαδή στην ουσία είναι Μηχανική Μάθηση, και χωρίζεται (όπως και η Μηχανική Μάθηση) σε SUPERVISED, UNSUPERVISED AND SEMI-SUPERVISED LEARNING

Εφαρμογές του deep learning στη βιοπληροφορική

* αναγνώριση/χαρτογράφηση ενζύμων και της δράσης τους με βάση την αλληλουχία των αμινοξέων τους, με χρήση πολύ-επίπεδων neural networks (multi-layer neural networks)
* πρόβλεψη γονιδιακής έκφρασης με χρήση RNN και CNN
* RNA-protein binding πρόβλεψη με χρήση CNN
* Biomedical image classification using transfer learning and ResNet
* Graph embedding for novel protein interaction prediction using GCN
* Biology image super-resolution using GAN
* High dimensional biological data embedding and generation with VAE
* sequence analysis 🡪 πρόβλεψη δράσης/συμπεριφοράς μη-κωδικοποιούμενων περιοχών (εσωνίων) στις αλληλουχίες του DNA/RNA (όπως και στα ΑΝΝs)
* πρόβλεψη δευτεροταγούς δομής πρωτεΐνης (όπως και στα ΑΝΝs) και πρόβλεψης αλληλεπίδρασης πρωτεινής με άλλες πρωτεινες και μακρομόρια (protein-protein molecular dynamics, protein-ligand docking, protein-dna docking)
* πρόβλεψη λειτουργικότητας/δράσης πρωτείνης
* διάγνωση ασθενειών μέσα από εφαρμογή deep learning σε αξονικές και μαγνητικές τομογραφίες (CT images, PET/PET-CT scan images, MRI images)

**Supervised Learning**

Αποτελείται από 3 δομικά στοιχεία:

1. Έναν γεννήτορα τυχαίων διανυσμάτων **Χ** προερχόμενων από μια φιξαρισμένη μη-γνωστή στατιστική κατανομή Ρ(**x**)
2. Έναν επιβλέποντα «supervisor» , ο οποίος (σαν στοιχείο-παράμετρος στο πρόβλημά μας) κάνει υπολογισμό μιας όχι διανυσματικής τιμής Υ για κάθε διανυσματική τιμή **Χ** σύμφωνα με μια υπό συνθήκη πιθανότητα Ρ(y|**x**)
3. Μια τάξη «μηχανών μάθησης» {f(**x**, θ): θ ∈ Θ}.

Τα ζεύγη (**X1**, Y1), . . . , (**Xn**, Yn), τα οποία είναι στατιστικώς ανεξάρτητα μεταξύ τους ως προς τις στατιστικές μεταβλητές **Χ,** αλλά και ομοιόμορφα κατανεμημένα, τα ονομάζουμε δεδομένα εκμάθησης (training data) προερχόμενα από την από κοινού κατανομή P(**x**, y) = P(y|**x**)P(**x**).

Στο supervised learning, το κεντρικό πρόβλημα που τίθεται είναι το εξής:

«Δοθέντων των δεδομένων εκμάθησης (όπως ορίστηκαν από πάνω), πώς μπορούμε να διαλέξουμε από την τάξη «μηχανών μάθησης» {f(**x**, θ): θ ∈ Θ} ένα κατάλληλο f (δηλαδή μια κατάλληλη «μηχανή μάθησης») η οποία να κάνει πρόβλεψη την απόκριση Υ του επιβλέποντα «supervisor» με τον βέλτιστο τρόπο»

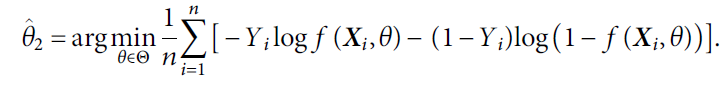
Εδώ, συναντάμε 2 υποπεριπτώσεις: το Υ να είναι συνεχές ή μη-συνεχές.

Όταν το Υ είναι συνεχές, το πρόβλημα μάθησης αποκαλείται πρόβλημα αναδρομής (regression problem), στο οποίο αυτό πρόβλημα αναζητάμε την καλύτερη τιμή της παραμέτρου θ η οποία να κάνει ελαχιστοποίηση την ακόλουθη συνάρτηση που λέγεται «quadratic loss function»



Όταν το Υ δεν είναι συνεχές , τότε το πρόβλημα μάθησης αποκαλείται πρόβλημα ταξινόμησης ή αναγνώρισης προτύπου (classification or pattern recognition problem)

Η αντίστοιχη εξίσωση για τη παράμετρο θ είναι η εξής :





Και έχουμε να προσθέσουμε ότι όταν , , τότε το θ1 μετατρέπεται στον κλασικό εκτιμητή ελαχίστων τετραγώνων στη γραμμική αναδρομή (least squares estimator in linear regression), και όταν τότε το θ2 μετατρέπεται στον εκτιμητή για τις συνιστώσες σε μια logistic regression.



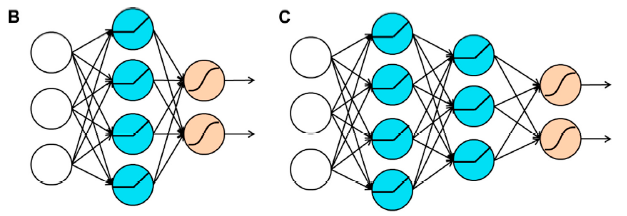
Μετά , έχουμε μια περιληπτική περιγραφή των εννοιών Neural Networks και Deep Neural Networks, που θα μας βοηθήσουν στην ανάλυση των άλλων δύο εννοιών **Unsupervised Learning** και **Semi-Supervised Learning**, αλλά και στην περιγραφή των εφαρμογών του Deep Learning σε genomics και bioinformatics και των εφαρμογών των unsupervised, supervised και semi-supervised learning κατ’ επέκταση.

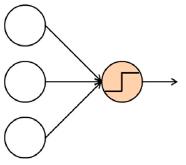
Neural Networks

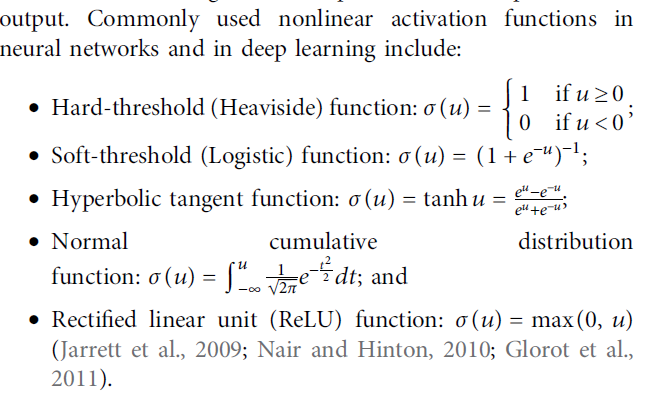
Είναι αλγόριθμοι που προσπαθούν να μιμηθούν τη συμπεριφορά του ανθρώπινου εγκέφαλου, αποτελούνται και αυτά από νευρώνες, και γι’ αυτό ονομάζονται και πολύ-επίπεδοι νευρώνες (multi-layer perceptrons) . Εάν χρησιμοποιήσουμε μη-γραμμικές συναρτήσεις ενεργοποίησης τότε το νευρωνικό δίκτυο δύναται να «πιάσει» μη-γραμμική σχέση μεταξύ των δεδομένων εισόδου και εξόδου, ενώ με γραμμικές τότε το νευρωνικό δίκτυο ανάγεται σε γραμμική ή λογιστική παλινδρόμηση. Σημαντικό χαρακτηριστικό ενός νευρωνικού δικτύου είναι το universal approximation theorem (Hornik et al., 1989), το οποίο έρχεται σε αντίθεση με τα πρόσφατα θεωρήματα των Györfi et al. (2006) και Shen et al. (2019) , δείχνοντας το κενό ανάμεσα στη θεωρία και την υλοποίηση στο συγκεκριμένο topic.

Δομή Νευρ.Δικτύου με 1 και 2 hidden layers

Δομή ενός perceptron





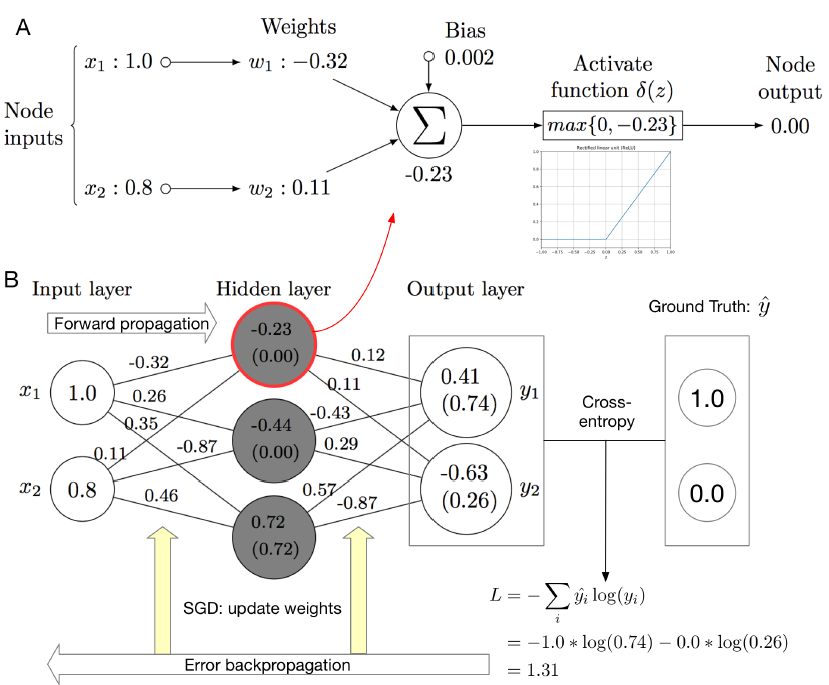


Shallow Neural Networks

Τα ρηχά νευρωνικά δίκτυα έχουν μόνο ένα hidden layer, και στην ουσία όλο το δίκτυο είναι μια mapping function (συνάρτηση απεικόνισης) (όπως φαίνεται παρακάτω στο σχήμα), και κατ’ επέκταση ο κάθε κόμβος του είναι μια mapping function.

Ο κάθε κόμβος του ΝΔ περιέχει δύο κομμάτια 🡪 έναν γραμμικό μετασχηματισμό w\*x (όπου w,x διανύσματα) και έναν μη-γραμμικό μετασχηματισμό δ(z).

Συγκεντρώνοντας πολλούς τέτοιους κόμβους δημιουργώντας ένα layer, και στη συνέχεια ενώνοντας δύο τέτοια layers, παράγεται το shallow neural network της παρακάτω εικόνας (το οποίο αναπαριστά μια μη-γραμμική συνάρτηση απεικόνισης (non-linear mapping function) μεγάλης πολυπλοκότητας).



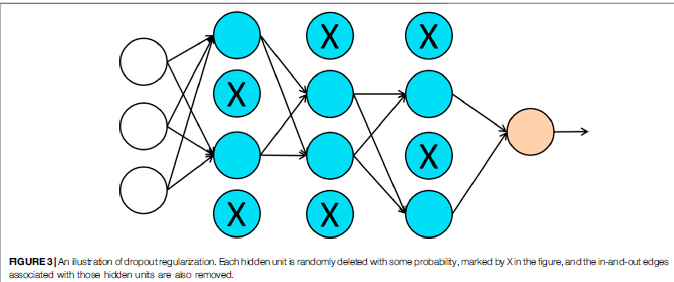
Ωστόσο, τα ρηχά νευρωνικά δίκτυα έχουν δύο σημαντικά μειονεκτήματα:

1. Ο αριθμός των παραμέτρων είναι εξαιρετικά αυξημένος 🡪 αν το shallow neural network έχει 2 hidden layers και το καθένα layer περιέχει Ν1 και Ν2 κόμβους αντίστοιχα, τότε στο τελικό ΝΔ θα έχουμε συνολικά N1\*N2 παραμέτρους, το οποίο μπορεί να προκαλέσει μεγάλη καθυστέρηση στην εκπαίδευση του νευρωνικού μας δικτύου.
2. Τα shallow neural networks δεν λαμβάνουν υπόψιν τις συσχετίσεις που μπορεί να έχουν τα δεδομένα εισόδου μας (πράγμα εξαιρετικά σύνηθες σε βιολογικά δεδομένα), και «θεωρούν» το κάθε δεδομένο εισόδου ως ανεξάρτητη οντότητα.

Deep Neural Networks

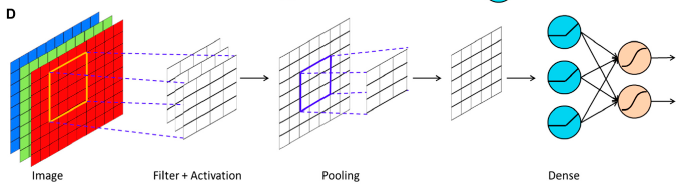
Είναι ένα νευρωνικό δίκτυο με παραπάνω από 1 hidden layer, και απαιτεί πολύ λιγότερο αριθμό κρυφών μονάδων για την εκμάθηση πολύπλοκων χαρακτηριστικών, ενώ μπορεί να χρειαστεί πολύ μεγαλύτερος αριθμός κρυφών μονάδων για ένα ρηχό νευρωνικό δίκτυο το οποίο έχει μαξ 1 hidden layer. Η πιο συνήθης μη-γραμμική συνάρτηση ενεργοποίησης είναι η ReLU, όπως φαίνεται στην από πάνω φωτογραφία.

Ένα γενικό πρόβλημα/δυσκολία στο deep learning είναι ότι με τον αυξανόμενο αριθμό εισόδων, οι περισσότεροι αλγόριθμοι μάθησης πρέπει να αντιμετωπίσουν το ζήτημα της υπερβολικής προσαρμογής (over-fitting issue), το οποίο ανάγεται (δηλαδή είναι ίδιο) με το bias-variance tradeoff στη στατιστική. Συνήθεις τρόποι αντιμετώπισης αυτής της δυσκολίας του deep learning είναι το regularization (κανονικοποίηση) και το dropout. Στο regularization, προστίθεται στο loss function μια τιμή ποινής (penalty term) ώστε να αντιμετωπιστεί το over-fitting issue, και η τιμή ποινής δύναται να αυξηθεί καθώς το μοντέλο προσπαθεί να προσεγγίσει με όσο το δυνατόν καλύτερο τρόπο τη πραγματική τιμή στην έξοδο. Στο dropout, διαγράφουμε τυχαία τις κρυφές μονάδες με συγκεκριμένη πιθανότητα και αφαιρούμε όλες τις ακμές εισόδου και εξόδου που σχετίζονται με αυτές τις κρυφές μονάδες.

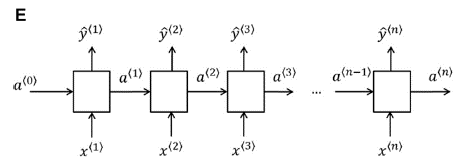


Δύο ακόμα διαδεδομένοι τύποι Νευρωνικών δικτύων

Είναι το συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network (CNN)), που χρησιμοποιείται για δομές δεδομένων πλέγματος (grid-like data structure) όπως εικόνες, και το αναδρομικό νευρωνικό δίκτυο (Recurrent Neural Network (RNN)) που χρησιμοποιείται για ακολουθιακά δεδομένα όπως DNA sequences.



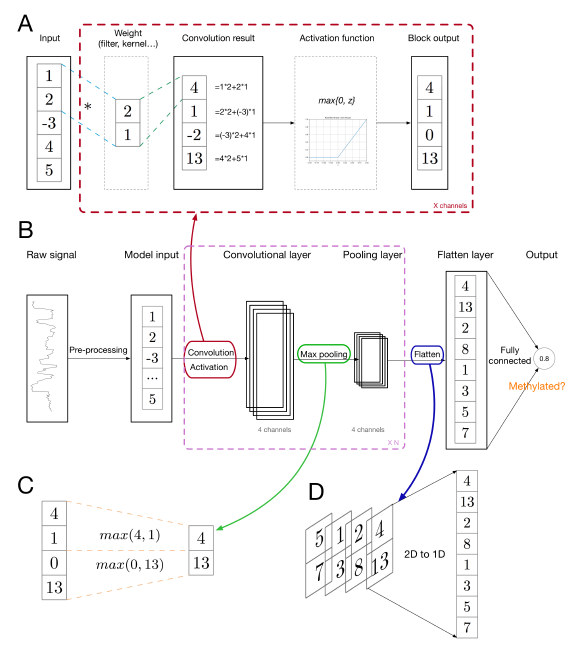
Δομή ενός CNN



Δομή ενός RNN

CNN

Κύριο χαρακτηριστικό του CNN είναι ότι κάνει συνελικτικούς υπολογισμούς (convolutional operations) αντί για υπολογισμούς μητρώων (matrix operations). Όντας ένα deep neural network, το CNN έχει κρυφά layers, τα οποία αποτελούνται από 2 μέρη/είδη. Το πρώτο είδος είναι το convolution layer (στρώμα συνέλιξης), και το δεύτερο ονομάζεται pooling layer. Σε κάθε συνελικτικό υπολογισμό εφαρμόζεται μια συνάρτηση ενεργοποίησης, παράγοντας έτσι το convolution layer.

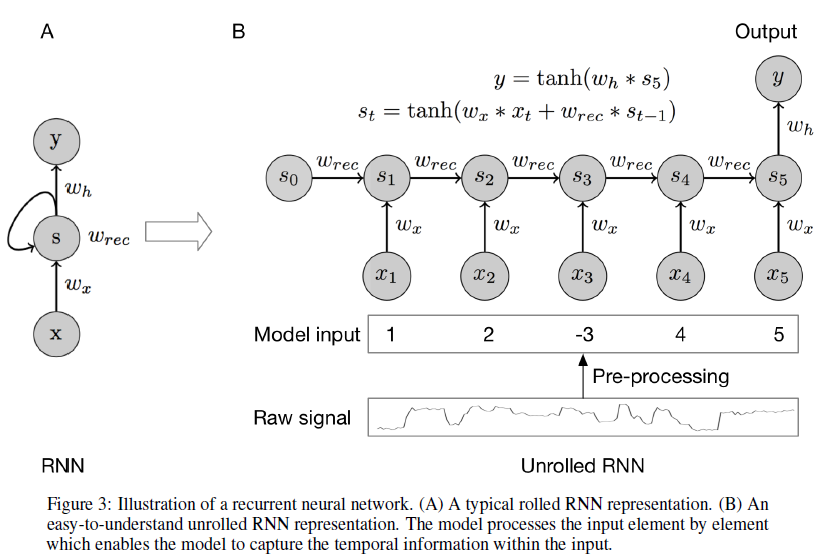


Γενική αρχιτεκτονική ενός CNN

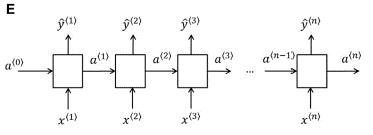
RNN

Ενώ το CNN χρησιμεύει στο να λάβουμε χωρική πληροφορία από τα δεδομένα μας, το RNN χρησιμοποιείται για να λάβουμε τη προσωρινή δυναμική συμπεριφορά των δεδομένων μας.

Τρία ειδικά χαρακτηριστικά ενός RNN είναι πρώτον ότι η είσοδος (πχ μια λέξη σε μια πρόταση) συνδυάζεται με την έξοδο του προηγούμενου hidden layer του RNN και ο συνδυασμός αυτός μπαίνει ως είσοδος στο επόμενο/διαδεχόμενο σκέλος-layer-neural network του RNN. Δεύτερον, το μήκος της εισόδου και της εξόδου του RNN μπορεί να είναι διαφορετικά, και τρίτον, οι παράμετροι του RNN διαδίδονται μεταξύ των neural network του, το οποίο βοηθάει στο να μη χρειάζεται μεγάλο ποσό παραμέτρων που πρέπει να εκτιμηθούν.

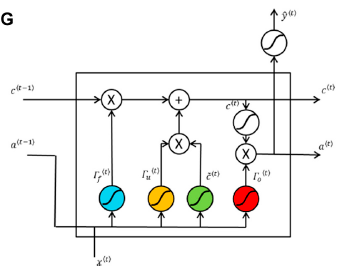


Γενική δομή ενός RNN



Δύο μειονεκτήματά του είναι ότι χρησιμοποιεί τις πληροφορίες νωρίτερα στην ακολουθία, το οποίο μπορεί να αντιμετωπιστεί χρησιμοποιώντας ένα αμφίδρομο RNN, και δεύτερον το ότι μπορεί να εγκλωβιστεί σε πρόβλημα κλίσης (gradient problem), που καθιστά δύσκολη την αποτύπωση εξαρτήσεων μεγάλης εμβέλειας στα δεδομένα μας. Το δεύτερο αυτό πρόβλημα λύνεται με δύο τροποποιήσεις του RNN : 1) το GRU (gated recurrent unit) και 2) το long short-term memory (LSTM) unit

Δομή του LMST



Forget gate

Update gate

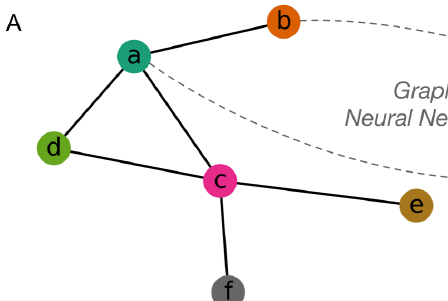
Graph neural networks

Τα graph neural networks διαχειρίζονται δεδομένα πλέγματος (network data), που

είναι συνήθης τύπος δεδομένων στη βιοπληροφορική, αλλά σύνθετος ωστόσο, αντίθετα με τα ακολουθιακά δεδομένα (DNA sequences), και ο λόγος είναι διότι στο graph neural network κάθε κόμβος μπορεί να έχει τυχαία σύνδεση με τους υπόλοιπους κόμβους του δικτύου (δλδ δε σημαίνει απαραίτητα ότι κάθε κόμβος του δικτύου συνδέεται με όλους τους υπόλοιπους κόμβους όπως συμβαίνει στα CNNs και τα RNNs).

Σημαντικός στόχος στα GNNs είναι να αποκρυπτογραφηθεί η πληροφορία που κρύβεται στην τοπολογία του δικτύου (δηλαδή γιατί είναι έτσι οι συνδέσεις μεταξύ των κόμβων), και παράλληλα με την πληροφορία κάθε κόμβου, να εξαχθεί ένα συνολικό συμπέρασμα για την πληροφορία που κρύβεται πίσω από όλη την αναπαράσταση του δικτύου.

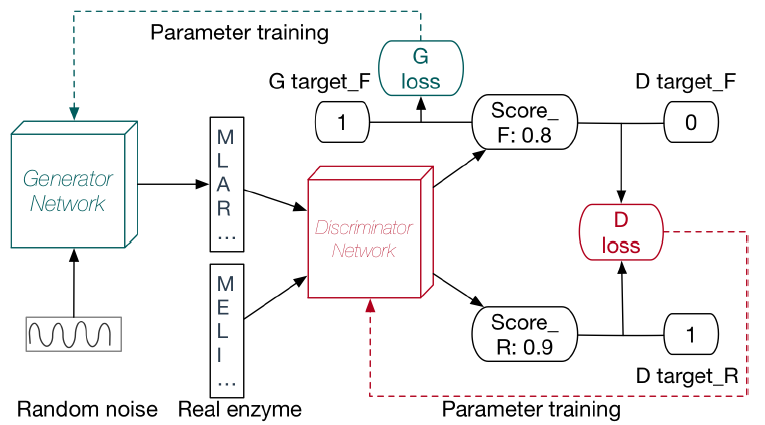
Πχ στην ακόλουθη φωτογραφία είναι ένα graph neural network και αναπαριστά μια αλληλεπίδραση πρωτεϊνών, όπου κάθε κόμβος είναι και μια πρωτεΐνη. Κάθε κόμβος (πρωτεΐνη) περιέχει, όπως είπαμε, και εσωτερική πληροφορία (δηλαδή αλληλουχία αμινοξέων και τρισδιάστατη στερεοδομή). Ο συνδυασμός και η επεξεργασία των δύο αυτών ειδών πληροφορίας (πληροφορίας συνδεσμολογίας δικτύου, και εσωτερική πληροφορία κάθε κόμβου (πρωτεΐνης)) είναι το κλειδί για να γίνει η εξόρυξη νέας γνώσης όσον αφορά το φαινόμενο-αλληλεπίδραση που αναπαρίσταται από τη συνολική δομή του GNN.



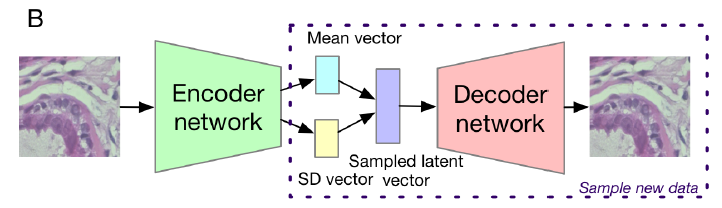
Generative models: GAN και VAE

Πολύ χρήσιμα όσον αφορά επεξεργασία βιοιατρικής εικόνας και in silico κατασκευή/σχεδιασμό πρωτεινών και φαρμάκων.

Χαρακτηριστικό του GAN είναι ότι αντί να εκπαιδεύει μόνο ένα νευρωνικό δίκτυο, το GAN εκπαιδεύει ένα ζεύγος δικτύων που ανταγωνίζονται μεταξύ τους, όπως φαίνεται στην παρακάτω φωτό, και το generator network είναι αυτό που παράγει νέα δεδομένα (πχ novel proteins/enzyme sequences).



Το VAE (Variational Auto Encoder) χρησιμοποιεί την έννοια του autoencoder που εισάγαμε νωρίτερα στην αναφορά μας, και περιέχει επιπροσθέτως ένα mean vector και ένα SD vector , και έχει τη μορφή της παρακάτω εικόνας.



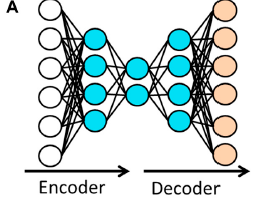
Εφαρμογές supervised learning στη γονιδιωματική

* Χρήση supervised machine learning αλγορίθμων όπως Support Vector Machine (SVM), Logistic Regression , Decision Tree , k-NN, Random Forests σε ζητήματα ανάλυσης γονιδιώματος και μελέτη αντισωμάτων (πχ SabDab by OPIG University of Oxford)
* DNA data sequencing
* genomic research and genetic association studies 🡪 πρόβλεψη της επιρροής/ρόλου των εσωνίων και των μη-κωδικοποιήσιμων περιοχών της αλληλουχίας DNA (πχ αν είναι υπεύθυνα για ασθένεις)🡪 DeapSea, DanQ
* gene expression inference 🡪 D-GEX 🡪 better than linear regression by 15,33%
* CNNs 🡪 πολύ καλά για data with special dependencies 🡪 DeepBind, DeepSea, Basset
* RNNs 🡪 χρησιμοποιούν matrix factorization (και όχι convolutional υπολογισμους όπως τα CNNs που αναφέρθηκε παραπάνω) 🡪 κατασκευή genotype and phenotype sequence prediction system 🡪 DeepNano , συνδυασμός με Particle Swarm optimization για αποκωδικοποίηση της γονιδιακής έκφρασης, ProLanGo για πρόβλεψη δράσης των πρωτεινών
* Gene-gene interactions 🡪 χρήση GENN (grammatical evolution neural networks), GPNN (genetic programming neural networks), Bayesian Neural Networks
* Μοντελοποίηση αλληλεπίδρασης και σχέσης μεταξύ γενετικών δεικτών
* Deep-learning based πρόβλεψη επιβίωσης μέσα από το προφίλ γενετικής έκφρασης 🡪 SurvivalNet, Cox-nnet, SALMON

**Unsupervised Learning**

Στην επιβλεπόμενη μάθηση είδαμε προηγουμένως ότι υπάρχει ένας «δάσκαλος» (επιβλέποντας «supervisor») ο οποίος επιβλέπει την απόδοση της μηχανής μάθησης μέσω κάποιας μέτρησης που ποσοτικοποιεί την απόκλιση. Αυτό δεν ισχύει στην μη-επιβλεπόμενη μάθηση, στην οποία μας ενδιαφέρει περισσότερο η συμπίεση δεδομένων εξάγοντας χρήσιμες πληροφορίες από τα δεδομένα εισόδου. Η διάσταση των εξαγόμενων χαρακτηριστικών είναι συνήθως πολύ μικρότερη από τη διάσταση των αρχικών δεδομένων εισόδου. Με αυτόν τον τρόπο, μπορούμε όχι μόνο να μειώσουμε το κόστος της αποθήκευσης των δεδομένων μας, αλλά και να κάνουμε τις μεταγενέστερες αναλύσεις πιο αποτελεσματικές.

Συνήθης τεχνική της μη-επιβλεπόμενης μάθησης είναι το PCA (Principle Component Analysis), και μέσα στο βαθύ νευρωνικό δίκτυο υπάρχει ένα κομμάτι του το οποίο είναι αντίστοιχο του PCA, το οποίο λέγεται autoencoder (αυτό-κωδικοποιητής). Στον αυτό-κωδικοποιητή γίνεται εξαγωγή σημαντικών χαρακτηριστικών από τα αρχικά (original) δεδομένα μας, και στη συνέχεια προκειμένου να αποφανθούμε εάν τα εξαγμένα χαρακτηριστικά αυτά αναπαριστούν τα αρχικά δεδομένα μας, κάνουμε επανακατασκευή των αρχικών αυτών δεδομένων μας από τα εξαγμένα χαρακτηριστικά και χρησιμοποιούμε τη διαφορά μεταξύ των ανακατασκευασμένων και αρχικών δεδομένων σαν οδηγό για να εκπαιδεύσουμε το δίκτυο μας.



Δομή ενός autoencoder

Επίσης, ενεργό ερευνητικό πεδίο πάνω στη μη-επιβλεπόμενη μάθηση είναι τα generative models. Στόχος τους είναι να μάθουν τη στατιστική κατανομή του μοντέλου απ’ τα δεδομένα έτσι ώστε να μπορούμε να παράξουμε νέα δεδομένα από την κατανομή αυτή.

Συνήθη χρησιμοποιούμενα generative models:

* Boltzmann machines (BM)
* Restricted Boltzmann machines (RBM)
* Deep Belief Networks (DBN)
* Deep Boltzmann machines (DBMs)
* Variational autoencoders (VAE)
* Generative adversarial networks (GAN)

Εφαρμογές un-supervised learning στη γονιδιωματική και γενετική

* Χρήση autoencoders για εκτίμηση τιμών που απουσιάζουν (missing values) πάνω σε μεταβολιτικά και γονιδιακής έκφρασης δεδομένα.
* Χρήση autoencoders για κατασκευή υπογραφών γονιδιακής έκφρασης (gene expression signatures) σύμφωνα με βιολογικές οδούς 🡪 ADAGE και GSEA (gene set enrichment analysis)
* Εφαρμογές και των Generative Models στη γονιδιωματική 🡪 VASC (deep variational autoencoder for single-cell RNA sequencing data ) 🡪 αναπτύχθηκε για να μοντελοποιήσει τα γεγονότα εγκατάλειψης (dropout, όπως είπαμε και πιο πάνω) και να βρει τις μη γραμμικές ιεραρχικές αναπαραστάσεις των δεδομένων
* Κατασκευή ενός GAN για προσομοίωση single-cell RNA-seq data.
* Κατασκευή συνθετικών gene sequences (γονιδιακών ακολουθιών) με επιθυμητές ιδιότητες, δημιουργία γονιδίων που κωδικοποιούν αντιμικροβιακά πεπτίδια καθώς και βελτιστοποίηση συνθετικών γονιδίων για τη δευτερογενή δομή των πεπτιδίων που προκύπτουν. 🡪 feedback GAN (FBGAN)
* Χρήση DBN (Deep Belief Neural Networks) για αρχικοποίηση παραμέτρων σε ένα βαθύ νευρωνικό δίκτυο για μελέτες ποσοτικής σχέσης δραστηριότητας δομής (QSAR)
* δημιουργία τεχνητών γονιδιώματων υψηλής ποιότητας, και τα αποτελέσματα είναι πολλά υποσχόμενα και το παραγόμενο τεχνητό γονιδίωμα μπορεί να κληρονομήσει συσχετίσεις γονότυπου-φαινοτύπου
* Κατηγοριοποίηση (clustering) κυτταρικών τύπων με χρήση αυτοκωδικοποιητών και με ενσωμάτωση τιμών έντασης εικονοστοιχείων με μετρήσεις γονιδιακής έκφρασης από κηλίδες σε ιστό 🡪 SpaCell
* εξαγωγή χαρακτηριστικών από εικονοστοιχεία εικόνας που δημιουργήθηκαν από την εικόνα μικροσκοπίας με χρωματισμό με αιματοξυλίνη και ηωσίνη που περιέχει πληροφορίες μορφολογίας ιστών 🡪 StLearn ((Pham et al., 2020))
* εφαρμογές πάνω σε Ιστολογία και μελέτη ιστών (Ιατρική) 🡪 SRT (Hu et al. (2021))
* συγκέντρωση πληροφοριών γονιδιακής έκφρασης από γειτονικά σημεία και στη συνέχεια ανίχνευση χωρικά μεταβλητών γονίδια με βάση τη συγκεντρωτική γονιδιακή έκφραση 🡪 SpaGCN (Li et al., 2020)

**Semi-Supervised Learning**

Είναι το ενδιάμεσο στάδιο ανάμεσα στην επιβλεπόμενη και μη-επιβλεπόμενη μάθηση, και ο στόχος της είναι η κατασκευή ενός εκμαθητή (learner) χρησιμοποιώντας τόσο τα επισημασμένα δεδομένα εκπαίδευσης (labeled training data) όσο και τα μη επισημασμένα δεδομένα για βελτιωμένη απόδοση, πράγμα που δεν ισχύει στην επιβλεπόμενη μάθηση όπου γίνεται χρήση μόνο των labeled data. Επίσης, παρατηρείται συνεχής αύξηση επίδοσης εάν χρησιμοποιήσουμε ημί-επιβλεπόμενες μεθόδους μάθησης βασισμένες σε νευρωνικά δίκτυα συγκριτικά με την επιβλεπόμενη μάθηση, στην οποία επιβλεπόμενη μάθηση κάθε δεδομένο θεωρείται ως «διδάσκων» (teacher) στη διαδικασία της μάθησης και καθορίζει, ουσιαστικά, το πόσο αποδοτική είναι η πρόβλεψη.

Η ημι-επιβλεπόμενη μάθηση διακρίνεται σε δύο κατηγορίες μεθοδων: τις επαγωγικές (inductive) μεθόδους και τις μεταγωγικές (transductive) μεθόδους

Στόχος των επαγωγικών μεθόδων είναι παρόμοιος με αυτόν της επιβλεπόμενης μάθησης, δηλαδή «Δοθέντων των δεδομένων εκμάθησης (όπως ορίστηκαν από πάνω), πώς μπορούμε να διαλέξουμε από την τάξη «μηχανών μάθησης» {f(**x**, θ): θ ∈ Θ} ένα κατάλληλο f (δηλαδή μια κατάλληλη «μηχανή μάθησης») η οποία να κάνει πρόβλεψη την απόκριση Υ του επιβλέποντα «supervisor» με τον βέλτιστο τρόπο» , ενώ στόχος των μεταγωγικών μεθόδων είναι να προβλέπουν αποκλειστικά ετικέτες/επισημάνσεις (labels) για τα μη επισημασμένα δεδομένα.

Εφαρμογές semi-supervised learning στη γονιδιωματική και γενετική και σε single cell RNA-seq data

* κατηγοριοποίηση (classification) και ανάλυση κυτταρικών τύπων με χρήση μονοκυτταρικών RNA-seq δεδομένων 🡪 scNym (Kimmel and Kelley, 2020)
* Μεγάλη βοήθεια στην ακριβή αναγνώριση/αποκωδικοποίηση κυτταρικών τύπων ώστε να ερμηνευτεί στη πορεία ο σχηματισμός καρκινικών συσσωματωμάτων και όγκων.
* Χρήση singlecell Variational Inference (scVI) και single-cell ANnotation using Variational Inference (scANVI) για πλήρη χαρακτηρισμό των στατιστικών κατανομών στα μονοκυτταρικά RNA-seq δεδομένα, για σχολιασμό τύπων κυττάρων 🡪 scSemiCluster (Chen et al., 2021), CALLR (Wei and Zhang, 2021), scNym (Kimmel and Kelley, 2021)
* scReClassify 🡪 κάνει χρήση PCA με στόχο τη μείωση διαστάσεων στα αρχικά single cell RNA-seq δεδομένα, και στη συνέχεια κάνει εφαρμογή μιας ημι-επιβλεπόμενης μεθόδου μάθησης για να επαναταξινομήσει τους λανθασμένους τύπους κυττάρων που προέκυψαν από λανθασμένη ανθρώπινη αξιολόγηση (με τη προυπόθεση ότι η λανθασμένη αυτή ανθρώπινη αξιολόγηση είναι σε σχετικά μικρό ποσοστό)
* cell type classifications in single cell RNA-seq data 🡪 SCINA algorithm (Zhang et al. (2019))
* cell type annotation methods 🡪 scSemiCluster (Chen et al., 2021) and CALLR (Wei and Zhang, 2021)

**Βιβλιογραφία:**

1. <https://microbenotes.com/bioinformatics-introduction-and-applications/>.
2. <https://en.wikipedia.org/wiki/Structural_biology>
3. Singh, H., Bioinformatics: Benefits to Mankind. *Int. J. Pharm. Tech. Res.,* 9, 4,242−248, 2016.
4. <https://www.biotecharticles.com/Bioinformatics-Article/Applications-of-Bioinformatics-3270.html>.
5. Rhee, S.Y., Dickerson, J., Xu, D., Bioinformatics and its applications in plant biology. *Annu. Rev. Plant Biol.*, 57, 335–360, 2006.
6. Lancashire, L.J., Lemetre, C., Ball, G.R., An introduction to artificial neural networks in bioinformatics—Application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings Bioinf.*, 10, 3, 315–329, 2009.
7. Wang, J.-Y., *Application of support vector machines in bioinformatics*, NationalTaiwan University, Taiwan, 2002.
8. <https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9>.
9. Polaka, I., Tom, I., Borisov, A., Decision tree classifiers in bioinformatics. *Appl. Comput. Syst.*, 42, 1, 118–123, 2010.
10. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearestneighbors-algorithm-6a6e71d01761>.
11. Qi, Y., Random forest for bioinformatics, in: *Ensemble machine learning*, pp. 307–323, Springer, Boston, MA, 2012.
12. Samundeeswari, E. and Saranya, P., An artificial neural network model for prediction of survival time of breast cancer dataset. *Int. J. Res. Eng. Appl. Sci.*, 6, 1, 161–168, 2016.
13. Narayanan, A., Keedwell, E.C., Gamalielsson, J., Tatineni, S., Single-layer artificial neural networks for gene expression analysis. *Neurocomputing*, 61, 217–240, 2004.
14. Hu, Y., Ashenayi, K., Veltri, R., O’Dowd, G., Miller, G., Hurst, R., Bonner, R., A comparison of neural network and fuzzy c-means methods in bladder cancer cell classification, in: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, vol. 6, IEEE, pp. 3461–3466, 1994.
15. Won, H.-H. and Cho, S.-B., Paired neural network with negatively correlatedfeatures for cancer classification in DNA gene expression profiles, in: *Proceedings of the International Joint Conference on Neural Networks, 2003*, vol. 3, pp. 1708–1713, IEEE, Portland, Oregon, 2003.
16. Thein, H.T.T. and Tun, K.M.M., An approach for breast cancer diagnosis classification using neural network. *Adv. Comput.*, 6, 1, 1, 2015.
17. Peterson, L.E. and Coleman, M.A., Machine learning-based receiver operatingcharacteristic (ROC) curves for crisp and fuzzy classification of DNAmicroarrays in cancer research. *Int. J. Approximate Reasoning*, 47, 1, 17–36, 2008.
18. Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Segovia Bucheli, V.,Rodrνguez-Sotelo, J.L., Jimιnez-Varσn, C.F., A comparative study of machine learning and deep learning algorithms to classify cancer types based onmicroarray gene expression data. *PeerJ Comput. Sci.*, 6, e270, 2020.
19. Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT press.
20. Wei, J.S., Greer, B.T., Westermann, F., Steinberg, S.M., Son, C.-G., Chen,Q.-R., Whiteford, C.C. *et al.*, Prediction of clinical outcome using geneexpression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Res.*, 64, 19, 6883–6891, 2004.
21. Cangelosi, D., Pelassa, S., Morini, M., Conte, M., Bosco, M.C., Eva, A.,Sementa, A.R., Varesio, L., Artificial neural network classifier predicts neuroblastomapatients’ outcome. *BMC Bioinf.*, 17, 12, 347, 2016.
22. . Nayeem, Md O. G., Ning Wan, M., Hasan, Md K., Prediction of disease levelusing multilayer perceptron of artificial neural network for patient monitoring.*Int. J. Soft Comput. Eng. (IJSCE)*, 5, 7−23, 2015.
23. Bordoloi, H. and Sarma, K.K., Protein structure prediction using multipleartificial neural network classifier, in: *Soft Computing Techniques in VisionScience*, pp. 137–146, Springer, Berlin, Heidelberg, 2012.
24. Shanthi, D., Sahoo, G., Saravanan, N., Input Feature Selection using HybridNeuro-Genetic Approach in the diagnosis of Stroke. *Int. J. Comput. Sci. Netw.Secur.,* 8, 12, 2008.
25. Vapnik, V. N. (1998). Statistical Learning Theory. 1 edition. New York: Wiley- Interscience.
26. LeCun, Y. (1989). “Generalization and Network Design Strategies,” in Connectionism in Perspective. Editors F. Fogelman-Soulié, L. Steels,

R. Pfeifer, and Z. Schreter (Elsevier Science). Citeseer.

1. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning Representations by Back-Propagating Errors. Cogn. Model. 5, 1.
2. Schuster, M., and Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. IEEE Trans. Signal Process. 45, 2673–2681. doi:10.1109/78.650093
3. Cho, K., Van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.,et al. (2014). “Learning Phrase Representations Using RNN Encoder-Decoderfor Statistical Machine Translation” in Conference on Empirical Methods inNatural Language Processing (EMNLP 2014). ArXiv Prepr. ArXiv14061078. doi:10.3115/v1/d14-1179
4. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958.
5. Wager, S., Wang, S., and Liang, P. S. (2013). “Dropout Training as AdaptivemRegularization,” in Advances in Neural Information Processing Systems, 351–359.
6. Mehmood, M.A., Sehar, U., Ahmad, N., Use of Bioinformatics Tools in Different Spheres of Life Sciences. *J. Data Mining Genomics Proteomics*, 5,1–13, 2014.
7. Singh, H., Bioinformatics: Benefits to Mankind. *Int. J. Pharm. Tech. Res.,* 9, 4,242−248, 2016
8. Rhee, S.Y., Dickerson, J., Xu, D., Bioinformatics and its applications in plant

biology. *Annu. Rev. Plant Biol.*, 57, 335–360, 2006.

1. Singh, H., Bioinformatics: Benefits to Mankind. *Int. J. Pharm. Tech. Res.,* 9, 4, 242−248, 2016.
2. . https://www.biotecharticles.com/Bioinformatics-Article/Applications-of-Bioinformatics-3270.html.
3. <https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-isimportant/>.
4. <https://www.analyticssteps.com/blogs/understanding-bioinformatics-application-machine-learning>.
5. Lausen, B., Bioinformatics and classification: The analysis of genome expressiondata, in: *Classification, Clustering, and Data Analysis*, pp. 455–461,Springer, Berlin, Heidelberg, 2002.