ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ

ΚΑΤΕΥΘΥΝΣΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ

Νικόλαος Μπεγέτης

ΑΜ: ΠΙΒ0111

Αναφορά προόδου

Θέμα

Επέλεξα να μελετήσω τα παθογόνα SNPs για σημαντικές ομάδες ασθενειών και να βρω συσχετίσεις στη μετάλλαξη των SNPs για αυτές τις ασθένειες. Αυτό το κάνω για κάθε μία ασθένεια συγκρίνοντάς την με τις υπόλοιπες ασθένειες.

Δεδομένα

Τις κατηγορίες των ασθενειών τις διάλεξα με προσοχή ώστε να διαφέρουν ριζικά μεταξύ τους και να μην αποτελούν η μία υποσύνολο της άλλης. Μία πηγή που βοήθησε αρκετά σε αυτή τη διαφοροποίηση και την επιβεβαίωσα από άλλες πηγές ώστε να χρησιμοποιήσω τις κατηγορίες ασθενειών που θα αναφέρω παρακάτω, είναι η http://en.wikipedia.org/wiki/Lists_of_diseases . Η λίστα των ασθενειών αποτελείται από τις παρακάνω ασθένειες:

1. Cancer types
2. Cutaneous conditions
3. Endocrine diseases
4. Eye diseases and disorders
5. Intestinal diseases
6. Infectious diseases
7. Communication disorders
8. Genetic disorders
9. Neurological disorders
10. Voice disorders
11. Vulvovaginal disorders
12. Mental illness

Τα δεδομένα μας για αυτούς τους τύπους τα πήραμε από τη βάση για SNPs στο NCBI και συγκεκριμένα από εδώ: http://www.ncbi.nlm.nih.gov/snp αναζητώντας τα SNPs με το όνομα της ασθένειας. Για να είμαστε όμως ακριβείς και να μην παραλείψουμε δεδομένα μιας και για παράδειγμα κάποιες ασθένειες μπορεί να δηλώνονται με παραπάνω από ένα

ονόματα, χρησιμοποιήσαμε τη βάση OMIM (http://www.omim.org/) που υπάρχει για τα SNPs η οποία συσχετίζει όλα τα ονόματα που μπορεί να υπάρχουν για μία ασθένεια και είναι ουσιαστικά ή επίθετα (π.χ. ένα παράδειγμα είναι ο καρκίνος. Έγινε συσχέτιση του cancer με τις λέξεις 1.carcinoma, 2.neoplasia και 3.tumor). Με τον παραπάνω τρόπο καταφέραμε και συλλέξαμε ένα άρτιο dataset από την dbSNP του NCBI. Τα επερωτήματα που θέσαμε εκεί ήταν τα εξής:

1. cancer OR ((carcinoma OR neoplasia OR tumor))
2. communication
3. cutaneous OR ((cutis OR dermal OR skin OR tegument))
4. endocrine
5. eye OR ((enophthalmos OR enophthalmia OR "deep-set eyes" OR "sunken eyes" OR exophthalmos OR exophthalmia OR proptosis OR "prominent eyes" OR globe OR "eye region"))
6. genetic disorders OR ((androgenetic OR hermaphrodite OR intersex OR pseudohermaphrodite OR "sex reversal" OR "disorder of sex development"))
7. infection OR (("urinary tract infection" OR cystitis OR pyelonephritis OR UTI))
8. intestinal
9. mental OR (("developmental delay" OR "intellectual disability" OR "intellectual impairment" OR "mental retardation" OR "mentally retarded" OR "cognitive delay"))
10. neurologic
11. voice
12. vuvlovaginal

Η dbSNP βάση μας έδωσε ένα αριθμό δηλωμένων SNPs για την κάθε κατηγορία, που ανταποκρίνεται σε SNPs που αφορούν κάποια ασθένεια σε όλους του οργανισμούς αλλά και μόνο στον οργανισμό του ανθρώπου. Εμείς τελικά αποφασίσαμε να εφαρμόσουμε την έρευνά μας μόνο στα SNPs που αφορούν τον άνθρωπο μιας και από ό,τι παρατηρήσαμε όταν ξεκινήσαμε να κάνουμε την αναζήτηση σε όλους τους οργανισμούς, τα SNPs που έχουν καταγραφεί για άλλους οργανισμούς και τα οποία δεν υπάρχουν στον οργανισμό του ανθρώπου είχαν σημαντική έλλειψη δεδομένων που δεν μας βοηθούσε στην εξαγωγή συμπερασμάτων και που χαλούσε την εκτέλεση του αλγορίθμου RVM αφού έπρεπε να συμπληρώσουμε αυτή την έλλειψη δεδομένων ως μία ξεχωριστή κατηγορία, την «έλλειψη δεδομένων» που όμως επηρέαζε τα αποτελέσματα. Παρακάτω παραθέτουμε των αριθμό των δηλωμένων SNPs που βρέθηκαν συνολικά για κάθε οργανισμό και συγκεκριμένα για τον ανθρώπινο οργανισμό.

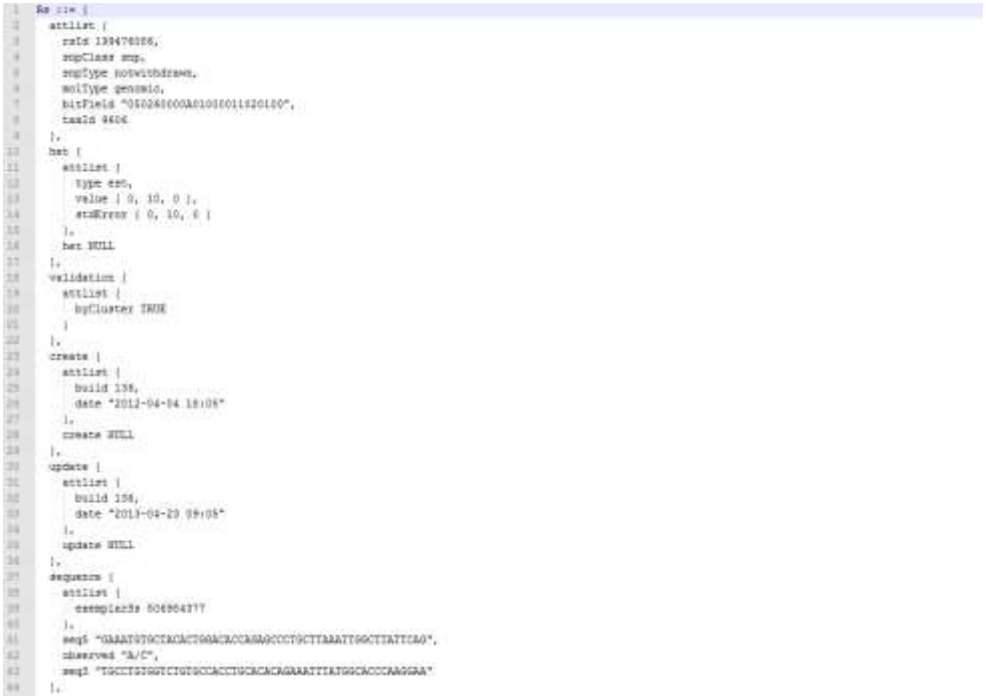
Πίνακας 1: Πλήθος δηλωμένων SNPs για κάθε ασθένεια που εξετάζουμε για όλους τους οργανισμούς και συγκεκριμένα για τον ανθρώπινο οργασνισμό

	All Organisms	Human
Cancer Types	424	217
Cutaneous Conditions	83	42
Endocrine Diseases	70	35
Eye Diseases and Disorders	399	56
Intestinal Diseases	39	20
Infectious Diseases	94	49
Communication Disorders	12	6
Genetic Disorders	6	3
Mental Illness	550	275
Neurologic Disorders	278	139
Voice Disorders	10	5
Vulvovaginal Disorders	2	1

Η βάση dbSNP του NCBI μπορούσε να μας δώσει όλα τα παραπάνω δεδομένα για κάθε κατηγορία είτε σε σύντομη περιγραφή των SNPs είτε σε αναλυτική μορφή. Για τη σύντομη μορφή τα διαθέσιμα format ήταν: 1.Brief, 2.FASTA, 3.FlatFile, 4.Summary και 5.Chromosome Report. Για την αναλυτική μορφή τα μόνα διαθέσιμα format ήταν το ASN.1 και το XML.

Για τη συγκεκριμένη εργασία, επειδή είχαμε ανάγκη για τη συλλογή όσων περισσότερων features μπορούσαμε να συλλέξουμε, διαλέξαμε να πάρουμε τα δεδομένα στην αναλυτική τους μορφή και να εξάγουμε από αυτά όσα περισσότερα features είναι δυνατό.

Η μορφή των row data αυτών των 2 τύπων ήταν αυτή που φαίνεται στις επόμενες 2 εικόνες:



Εικόνα 1: Μορφή τύπου ASN.1

Εικόνα 2: Μορφή XML

Όπως φαίνεται από τη μορφή των 2 αυτών τύπων είναι ευκολότερο να αναλύσουμε το πρώτο format με μία scripting γλώσσα όπως η BioPerl, ενώ το δεύτερο format είναι ευκολότερο να το αναλύσουμε με μία γλώσσα αντικειμενοστραφούς προγραμματισμού όπως η BioJava. Τελικά εμείς επιλέξαμε τη δεύτερη μέθοδο, γιατί αν και είναι δυσκολότερη η συγγραφή της, η BioJava είναι γρηγορότερη από την BioPerl και σε μεγάλα δεδομένα όπως τα αποτελέσματα των ασθενειών που κατεβάσαμε είναι καλύτερο το πρόγραμμά μας να μας εξάγει τα αποτελέσματα το συντομότερο δυνατό.

<u>Δεδομένα XML → Χαρακτηριστικά για την RVM μηχανή</u>

Με βάση όσα περιγράψαμε παραπάνω κατασκευάσαμε ένα project σε BioJava (περιλαμβάνεται στο παραδοτέο με όνομα “SeqAlignmentBiojava”) το οποίο εξάγει όλη τη χρήσιμη πληροφορία από το XML και την μετατρέπει ταυτόχρονα στην κατάλληλη είσοδο για την RVM μηχανή. Τα αποτελέσματα μετά από το parsing των XML αρχείων τα αποθηκεύσαμε στο φάκελο “SeqAlignmentBiojava output” του RVM project που μας δόθηκε στην τάξη και υλοποιεί τον αλγόριθμο του RVM. Τα αποτελέσματα έχουν όλα ονομαστεί με το όνομα της ασθένειας ακολουθούμενο από τη λέξη output (cancer_output.txt). Επίσης όλα τα αποτελέσματα έχουν ελεγχθεί και εξεταστεί από άλλο ένα δεύτερο πρόγραμμα που συμπεριλαμβάνεται στο ίδιο project και όλα έχουν τον ίδιο αριθμό χαρακτηριστικών (απαραίτητη προϋπόθεση για το RVM), ενώ χρησιμοποιήθηκε ένα τρίτο πρόγραμμα και αυτό στο ίδιο project το οποίο συνενώνει τα positive αποτελέσματα με τα negative σε ένα αρχείο αναμεμειγμένα ομοιόμορφα και τα οποία αργότερα χρησιμοποιήθηκαν για το training και testing στο RVM (θα μιλήσουμε για αυτά στη συνέχεια που θα περιγράψουμε τα αποτελέσματα του RVM).

Τα χαρακτηριστικά που χρησιμοποιήσαμε για την εύρεση της σημαντικότητάς τους στο διαχωρισμό 2 κλάσεων όπου η μία θα έχει τα positive data και η άλλη τα negative είναι τα εξής:

1. το χαρακτηριστικό της positive η negative κλάσης
2. rsId,
3. snpClass
4. snpType
5. molType
6. bitField
7. taxid
8. subSnpClass
9. orient
10. strand
11. molType
12. methodClass
13. validated
    - <u>for the forward sequence</u>
14. seq5conserved

15. a%
16. t%
17. g%
18. c%
19. g/c%
20. a/c%
21. a/g%
22. a/t%
23. g/t%
24. c/t%
25. ratio_a_t%
26. ratio_a_g%
27. ratio_a_c%
28. ratio_t_g%
29. ratio_t_c%
30. ratio_g_c%
31. ratio_a/t_g/c%
32. ratio_a/c_g/t%
33. ratio_a/g_c/t%
34. ratio_a_g/t%
35. ratio_a_g/c%
36. ratio_a_c/t%
37. ratio_t_a/g%
38. ratio_t_g/c%
39. ratio_t_a/c%
40. ratio_g_a/t%
41. ratio_g_a/c%
42. ratio_g_c/t%
43. ratio_c_a/t%
44. ratio_c_a/g%
45. ratio_c_g/t%
46. aa%
47. at%
48. ag%
49. ac%
50. tt%
51. tg%
52. tc%
53. gg%
54. gc%
55. cc%
56. aaa%
57. aat%
58. aag%
59. aac%
60. tta%
61. ttt%
62. ttg%
63. ttc%
64. gga%
65. ggt%
66. ggg%
67. ggc%
68. cca%
69. cct%
70. ccg%
71. ccc%
72. Observed Mutation
    - for the reverse sequence
73. seq3conserved
74. a%
75. t%
76. g%
77. c%
78. g/c%
79. a/c%
80. a/g%
81. a/t%
82. g/t%
83. c/t%
84. ratio_a_t%
85. ratio_a_g%
86. ratio_a_c%
87. ratio_t_g%
88. ratio_t_c%
89. ratio_g_c%
90. ratio_a/t_g/c%
91. ratio_a/c_g/t%
92. ratio_a/g_c/t%
93. ratio_a_g/t%
94. ratio_a_g/c%
95. ratio_a_c/t%
96. ratio_t_a/g%
97. ratio_t_g/c%
98. ratio_t_a/c%
99. ratio_g_a/t%
100. ratio_g_a/c%
101. ratio_g_c/t%
102. ratio_c_a/t%
103. ratio_c_a/g%
104. ratio_c_g/t%
105. aa%
106. at%
107. ag%
108. ac%
109. tt%
110. tg%
111. tc%
112. gg%
113. gc%
114. cc%
115. aaa%
116. aat%
117. aag%
118. aac%
119. tta%
120. ttt%
121. ttg%
122. ttc%
123. gga%
124. ggt%
125. ggg%
126. ggc%
127. cca%
128. cct%
129. ccg%
130. ccc%
131. A-T,G-C watson bonds %
132. A-C,G-T purine-pyrimidine bonds %
133. A-G,C-T pyrimidine-pyrimidine/purine-purine bonds %
134. A-A,T-T,G-G,C-C pyrimidine-pyrimidine/purine-purine bonds %
135. seq5pyrimidineSum
136. seq5purineSum
137. seq3pyrimidineSum
138. seq3purineSum
139. groupLabel
140. componentType

141. accession
142. chromosome
143. chromosome start
144. chromosome end
145. length (start-end)
146. orientation in chromosome
147. gi
148. groupTerm
149. contigLabel
150. locType
151. alnQuality
152. orient
153. refAllele
154. geneId
155. symbol
156. mrnaVer
157. protVer
158. fxnClass
159. readingFrame
160. allele
161. residue
162. aaPosition
163. chromCount
164. placedContigCount
165. unplacedContigCount
166. seqLocCount
167. hapCount
168. gi
169. accession
170. locType
171. alnQuality
172. orient
173. refAllele
174. protAcc
175. protGi
176. protLoc
177. protResidue
178. rsResidue
179. structGi
180. structLoc
181. structResidue
182. hgvs
183. AlleleOrigin

Όλα τα παραπάνω μπορείτε να τα βρείτε και στο αρχείο “rvm_features.txt” όπου δίνονται και κάποιες επιπλέον εξηγήσεις μέσω παραπομπών στον ιστό, γίνεται καλυτερη αντιστοίχιση όσων αφορά την κατανόηση των features με το XML αρχείο καθώς επίσης γίνεται και σχολιασμός στη μεθοδολογία που χρησιμοποιήσαμε για την μετατροπή των row data σε features (π.χ. το forward μετατράπηκε σε 1, ενώ το reverse σε 0).

RVM

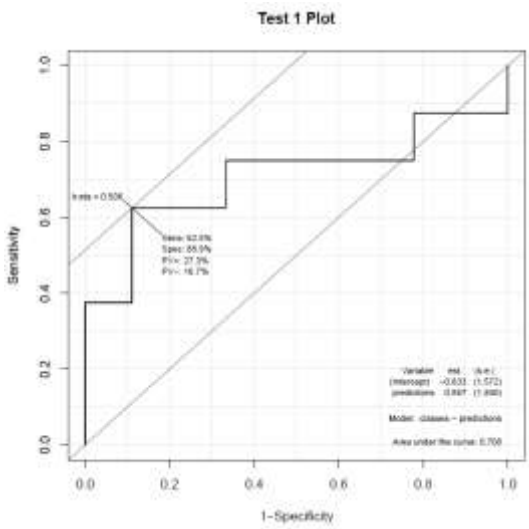
Έχοντας πλέον φτιάξει τα features σε μορφή που τα παίρνει το υλοποιημένο πρόγραμμα του RVM χωρίσαμε συγκρίναμε κάθε ασθένεια με κάθε άλλη. Δηλαδή, δώσαμε για παράδειγμα την ασθένεια του καρκίνου ως το positive set μας και ως negative δώσαμε ξεχωριστά κάθε άλλη από τις 11 ασθένειες. Αποφασίσαμε να το κάνουμε έτσι και να μην ορίσουμε ένα negative set με τυχαία δείγματα από τις άλλες 11 ασθένειες γιατί θέλαμε να βρούμε και τυχόν ομοιότητες στα SNPs μεταξύ των ασθενειών ανά δύο.

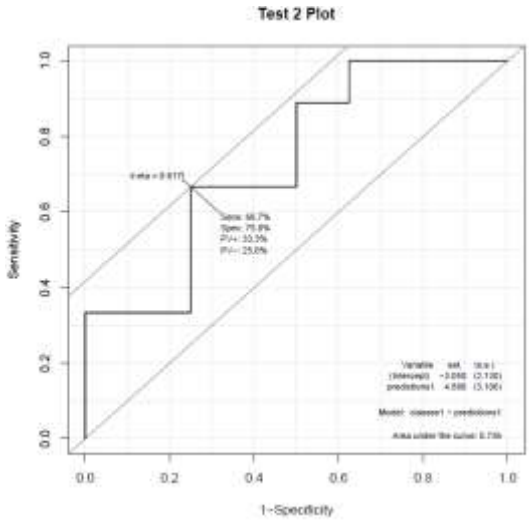
Όπως περιγράψαμε παραπάνω στο τρίτο πρόγραμμα που είναι υλοποιημένο στο SeqAlignmentBiojava κάναμε τις συνενώσεις των δεδομένων μεταξύ κάθε δύο ασθενειών και αυτές αποτελούσαν όλα το dataset που δώσαμε για το πρώτο training της μηχανής RVM πριν το 5-fold cross validation. Τα αρχεία αυτά αποτελούνται από ισάριθμα δεδομένα και από τις 2 κλάσεις που εναλλάσσονται εναλλάξ εντός του αρχείου. Μπορείτε να τα βρείτε στο φάκελο “RVM_BioJava/train” με ονόματα που περιέχουν τα 2 ονόματα των ασθενειών που χρησιμοποιήθηκαν για positive και negative data κάθε φορά ακολουθούμενα από τη λέξη train (π.χ. “cancer_endocrine_train”). Κάθε φορά χρειάστηκε τα αλλάζουμε το configuration input file του RVM_project για να κάνουμε εναλλάξ το training και το testing, όπως ενδείκνυται από τη βιβλιογραφία. Συνεπώς εμείς κάναμε πρώτα ένα training με όλο το set και μετά κάναμε το 5-fold cross-validation. Μετά από το πρώτο training της μηχανής ακολούθησε το 5-fold cross-validation και τα 5 train sets που κάθε φορά αποτελούν τα 4/5 του συνολικού δεδομένου μαζί με τα 5 test sets που αποτελούν το 1/5 του συνολικού

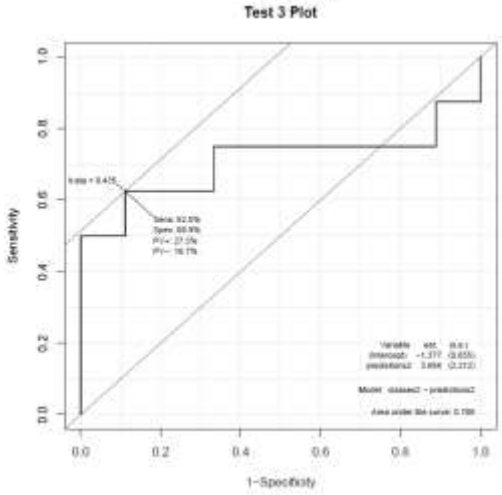
δεδομένου μπορείτε και αυτά να τα βρείτε στους φακέλους “RVM_BioJava/train” και “RVM_BioJava/test” μαζί με τα αντίστοιχα ονόματα του x-fold cross-validations για x από 1 μέχρι 5. Το RVM με τη σειρά του, κάθε φορά, εξήγαγε 2 αρχεία και εκτύπωνε στην οθόνη κάποια βάρη τα οποία τα αποθηκεύσαμε σε αρχεία κειμένου με τα αντίστοιχα οικεία ονόματα στο φάκελο “RVM_BioJava/run results”. Από τα δύο αρχεία του RVM, το πρώτο τα βάρη τα οποία έβγαιναν μετά από κάθε training και δίνονταν στο testing αρχείο, ενώ το δεύτερο εξαγόταν μετά από κάθε testing και αποτελούσε τα βάρη που δίνονταν στα πιο σημαντικά χαρακτηριστικά που συνιστούν το διαχωρισμό των κλάσεων.

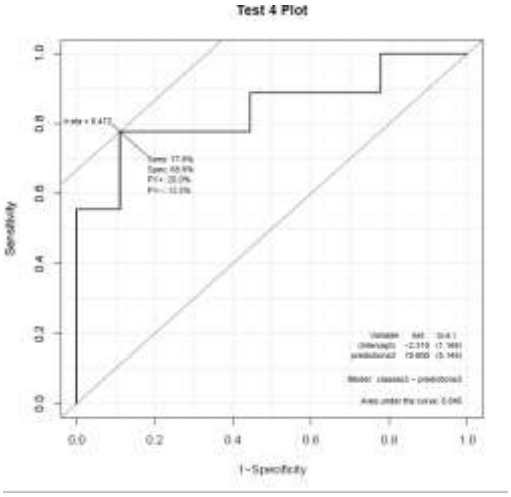
Σχήματα και Συμπεράσματα

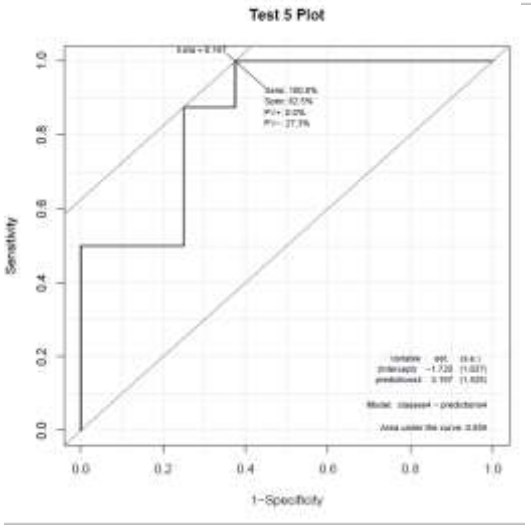
Μας ζητήθηκε να κάνουμε ROC-curves και Radar-Plot για κάθε ένα αποτέλεσμα μετά από το testing για κάθε ασθένεια. Επειδή στην περίπτωσή μας αυτά τα plots είναι πάρα πολλά (>100) παραθέτουμε ενδεικτικά σε αυτή την αναφορά (επόμενη σελίδα) τα 5 test plots της ασθένειας του Καρκίνου δίνοντας του ως negative set δεδομένα από γενετικές διαταράξεις. Δεν είναι τυχαίο το παράδειγμα, που παραθέτουμε. Στο συγκεκριμένο παράδειγμα τα plots δεν είναι καλοσχηματισμένα και φαίνεται ότι ο διαχωρισμός δεν έχει γίνει πολύ σωστά. Αυτό σημαίνει ότι πιθανότατα οι 2 ασθένειες έχουν αρκετά κοινά SNPs που επηρεάζουν τον ανθρώπινο οργανισμό. Αντίθετα, σε άλλες ασθένειες που χρησιμοποιήσαμε ως negative set με τον καρκίνο ο διαχωρισμός ήταν αρκετά πιο σαφής και σε ορισμένες περιπτώσεις απόλυτος. Όλα τα αποτελέσματα είναι αποθηκευμένα σε μορφή pdf για κάθε ασθένεια με όλες τις άλλες ασθένειες ως negative sets στο φάκελο “some ROC”.











ROC curves of Cancer as a positive dataset and Genetics Disorders as a negative dataset