

Maskininlärning med disjunktiv normalform

Anton Björklund

Kandidatavhandling

Åbo Akademi

Fakulteten för naturvetenskaper och teknik

Datavetenskap

Handledare: Ion Petre och Sepinoud Azimi

10.04.2016

Referat

I avhandlingen presenteras en maskininlärningsmetod som söker signaturer för kluster i disjunktiv normalform. Användningen av disjunktiv normalform resulterar i signaturer som är läsliga för en människa, vilket är den främsta fördelen med den här metoden. För att ge perspektiv på maskininlärning innehåller avhandlingen även korta beskrivningar på några alternativa maskininlärningsmetoder.

Förutom en grundlig genomgång av inlärningsmetoden och dess beståndsdelar beskrivs också det material som använts för att utvärdera metoden. Inlärningsmaterialet består av genuttrycksdata för några olika sjukdomar, bröstcancer, alzheimer och huvud- och nackcancer. För detta material söker inlärningsmetoden signaturer för olika sjukdomstillstånd.

Med resultaten för de olika datamängderna kan man utreda inlärningsmetodens förmåga att skapa läsliga och väldefinierade signaturer för alla kluster. Dessutom är också metodens snabbhet och begränsningar också av intresse för slutledningen.

Innehållsförteckning

Referat	1
Innehållsförteckning	2
1 Inledning	5
2 Bakgrund	6
2.1 Alternativ	6
2.1.1 Stödvektormaskiner	6
2.1.2 Beslutsträd	6
2.1.3 Artificiella neuronnät	7
2.2 Läslighet	7
2.2.1 Disjunktiv normalform	8
3 Metodik	9
3.1 Diskretisera data	9
3.1.1 Mediandiskretisering	9
3.1.2 Klusterdiskretisering	10
3.2 Signaturinlärning	11
3.2.1 Minimera antalet egenskaper	12
3.2.2 Signaturminimering	12
3.2.3 Espresso	13
4 Data	16
4.1 Gener	16

4.2 Genuttryck	17
4.3 Material	17
5 Inläring	19
5.1 Implementation	19
5.2 Resultat	19
5.3 Test	21
6 Avslutning	24
6.1 Utvärdering	24
6.1.1 Signaturkvalitet	24
6.1.2 Läslighet	24
6.1.3 Prestanda	25
6.2 Begränsningar	25
6.3 Slutsatser	25
Referenser	27
Bilagor	29
Bilaga 1: Signaturer för bröstcancerkluster	29
Normal	29
Non-basal-like cancer	29
BRCA1-associated cancer	30
Basal-like cancer	30
Bilaga 2: Signaturer för Alzheimerkluster	31
Alzheimer	31

Frisk	35
Bilaga 3: Huvud- och nackcancersignaturer	40
Cancer fas 0	40
Cancer fas 1	40
Cancer fas 2	40
Cancer fas 3	42
Cancer fas 4a	43
Cancer fas 4b	48
Cancer fas 4c	48
Cancer fas okänd	48

1 Inledning

I den här avhandlingen kommer jag presentera en maskininlärningsmetod som söker signaturer för kluster i disjunktiv normalform. Maskininlärning är något man kan tillämpa då datamängder blir för stora, för komplexa eller på annat sätt opraktiska för en människa att bearbeta. I dagens läge då datalagring är billigt skapas används maskininlärning ofta för att hitta och utnyttja kända och okända egenskaper och samband i allehanda ihopsamlade datamängder.

Det finns många olika metoder för att utvinna dold kunskap. Dessa metoder är anpassade för olika ändamål och situationer. Inlärningsmetoden som presenteras i den här avhandlingen fokuserar på att skapa resultat som är möjliga för en människa att förstå. Att en människa kan läsa resultatet gör att informationen som metoden upptäcker kan användas för vidare kunskap, till exempel som bas för vidare forskning. Läsligheten åstadkoms genom användandet av en boolesk logisk normalform, disjunktiv normalform.

Genuttrycksdata för tre olika sjukdomar, bröstcancer, alzheimer och huvud- och nackcancer, har använts för att testa inlärningsmetoden. Resultatet som fås från inlärningsmetoden blir då signaturer bestående av aktiva och inaktiva gener. Med resultatet från inläringen och testerna kan man konstatera att inlärningsmetoden är väldigt nischad och kräver tillräckligt mycket förhandskunskap för att fungera bra.

2 Bakgrund

Empiriska resultat har länge varit viktig inom vetenskaper för att komma fram till ny kunskap. I dagens läge har vi möjlighet att både samla och spara mera information än någonsin tidigare. I detta kapitel kommer jag inleda med ett urval av metoder för att åstadkomma detta lärande genom modern teknologi. Därefter flyttas fokuset till den maskininlärningsmetod som avhandlingen kretsar kring. I avsnitt 2.2 förklaras inlärningsmetodens grund som som en del av kunskapsskapandet.

2.1 Alternativ

Det finns många olika implementationer av maskininläring. Nedan kommer jag presentera några maskininlärningsmetoder som kan tillämpas på liknande problem som avhandlingens huvudmetod. Detta ger kontext till inlärningsmetoden.

2.1.1 Stödvektormaskiner

Stödvektormaskiner (SVM, Support Vector Machine) lär sig klassificera två kluster. Detta sker genom att först skapa vektorer för klustren baserat på egenskapernas värden. Sedan söks ett plan som delar vektorerna i de två klustren. Stödvektormaskiner har visats kunna klassificera okända fall med relativt hög precision samt klara av brus, det vill säga felaktiga klassificeringar i inlärningsdatan. Det plan en SVM lär sig och baserar sina beslut på är dock väldigt svår att representera på ett sätt som gör det möjligt för en människa att ta till sig den inlärd informationen [4].

2.1.2 Beslutsträd

Beslutsträd är sökträd som byggs upp av förgreningar som sker då egenskaper förgrenar sig i möjliga värden. Så länge träden inte blir alldeles för stora så är det ganska naturligt att läsa och förstå ett sökträd [3]. Som förklaras i avsnitt 2.2.1 kan resultatet från avhandlingens inlärningsmetod

presenteras i ett beslutsträd. Dock är metoden för att komma fram till resultatet är annorlunda.

2.1.3 Artificiella neuronnät

Artificiella neuronnät (ANN, Artificial Neural Network) är inspirerade av biologiska hjärnor. Neuronnät består av noder, neuroner, som är ihopkopplade till ett nätverk. Inläring sker genom att noder, kopplingar och kopplingars viktningar förändras. ANN har visat sig vara väldigt kraftfulla i många maskininläringssituationer. Även om iden bakom neuronnät baseras på hjärnor är dock kunskap som finns i nätverken svår att tyda för en människa [5].

2.2 Läslighet

De metoder som presenterades i föregående avsnitt tenderar fokusera på att hitta samband med största möjliga precision. Med precision avses hur bra metoden kan klassificera nya fall. Den metod som undersöks i denna avhandling lägger större fokus på att skapa resultat som kan läsas av en människa. Inlärningsmetodens förmåga att definiera kluster är naturligtvis också viktig.

Ett läsligt resultat kan vara önskvärt av flera olika orsaker. Om materialet som undersöks är välkänt kan en expert verifiera resultatet som fås. I fall det undersökta ämnet inte är tillräckligt känt kan det maskininlärd resultatet användas som grund för fortsatt forskning och kunskap. En människa är nödvändig i det här fallet eftersom maskinen endast arbetar med abstrakt data. Med ett läsligt resultat kan då en expert sätta in resultatet i en större sammanhang.

För att ett resultat ska vara läsligt behövs ett gemensamt sätt att kommunicera på. För den undersökta inlärningsmetoden används boolesk logik, som väldigt naturligt för en dator och dessutom är möjligt att förstå för en människa. Mera specifikt används en normalform i boolesk logik, disjunktiv normalform (hädanefter DNF). I nästa avsnitt förklaras vad DNF är

och dess styrka. Metodens användning av DNF i både sökning av signaturer och resultat presenteras i kapitel 3.

2.2.1 Disjunktiv normalform

DNF är en boolesk-logisk normalform. En normalform är ett strukturerat sätt att ställa upp en logisk formel på. I korthet innebär DNF en disjunktion av konjunktioner av atomära satser. Det betyder att variabler och negerade variabler sammanbinds till konjunktioner (satser med och-operatorer). Dessa konjunktioner sammanbinds disjunktivt till fulla uttryck (disjunktioner, satser med eller-operatorer). Tabell 2.1 visar ett exempel på ett uttryck i DNF.

Det är möjligt att bevisa att det för varje logisk formel existerar en ekvivalent formel i DNF (bevisas inte här). Detta gör att DNF har den fulla uttrycksförmågan och precisionen av boolesk logik. Dessutom kan DNF lätt omvandlas till och från en sanningstabell, vilket är väldigt användbart eftersom inlärningsdata ofta är i form av tabeller. Se tabell 2.1 för ett exempel på hur en sanningstabell kan omvandlas till ett uttryck i DNF [6].

A	Sant	Falskt	$(A \wedge \neg B \wedge \neg C \wedge D) \vee (\neg A \wedge B \wedge C \wedge D)$
B	Falskt	Sant	
C	Falskt	Sant	
D	Sant	Sant	

Tabell 2.1: Exempel på en sanningstabell representerad i DNF

Det är DNF:s starkt hierarkiska struktur som gör DNF förhållandevis läsligt. Strukturen är också till fördel då uttryck presenteras på andra sätt än i en formel. Exempel på visualiseringar är beslutsträd och sanningstabeller, något som syns i tabell 2.1. Visualiseringar kan ge en mera intuitiv förståelse av resultatet. Detta är viktigt då mänskligt förståelse av resultatet är en viktig kvalitet hos den maskininlärningsmetoden [15].

3 Metodik

Detta kapitel beskriver i detalj hur inlärningsprocessen går till. Metoden består av några mindre steg. Allt börjar med materialet som undersöks. Datan måste omvandlas till booleska värden för att kunna beskrivas i DNF. Denna transformation går igenom i avsnitt 3.1. Med alla förberedelser gjorda kan själva inläringen börja. Inläringen består av att hitta och optimera signaturer. Avsnitt 3.2 presenterar allt som ingår i inläringen.

3.1 Diskretisera data

Inlärningsalgoritmen och DNF använder sig av diskreta booleska värden. Innan DNF-signaturerna kan hittas måste således alla kontinuerliga värden omvandlas till booleska värden. Om klustrens storlekar är balanserade och tillräckligt stora kan egenskapernas medianer användas som bas för diskretiseringen. Om klustren inte är jämt fördelade eller om egenskapernas värden förväntas vara i två distinkta grupper kan en klustringsmetod användas för diskretisering.

3.1.1 Mediandiskretisering

Medianen är det värde som delar en mängd i två lika stora delar. För varje egenskap beräknas en median som används som tröskelvärde för diskretiseringen. Alla värden större än medianen byts ut mot 1 och alla mindre byts ut mot 0. Tabell 3.1 visualiserar hur detta går till.

Problem med användningen av medianer uppstår då värdena för en egenskap inte är jämt fördelade mellan högre och lägre värden. Då kommer en del värden att diskretiseras fel eftersom hälften måste ersättas med 1 och hälften med 0. Medianmetoden drabbas dock inte av problem som beror på hur värdena är fördelade sinsemellan över utfallsspektrumet, vilket klustringsmetoden i nästa avsnitt gör.

	Ursprunglig Data				Median	Diskretiserad Data			
	Sampl 1	Sampl 2	Sampl 3	Sampl 4		Sampl 1	Sampl 2	Sampl 3	Sampl 4
Egenskap 1	0,5	1,0	2,5	3,0	1,75	0	0	1	1
Egenskap 2	4,5	3,2	6,0	2,0	3,85	1	0	1	0
Egenskap 3	0,2	0,3	0,2	0,4	0,25	0	1	0	1
Egenskap 4	0,8	5,6	6,7	0,7	3,40	0	1	1	0

Tabell 3.1: Exempel på hur data diskretiseras med medianen som tröskelvärde

3.1.2 Klusterdiskretisering

Här används en tillämpad k-medelvärdes klustringsalgoritm (k-means clustering). Slutresultatet är att alla värden ändras till antingen 1 eller 0. Detta betyder att det söks två kluster. Observera att de kluster som används här är specifik för varje egenskap och har inget att göra med klustren i datamängden som samplen tillhör. Utgångspunkterna för klustren baseras på det minsta och det största värdet.

Algoritmen bygger på en loop. Först sätts alla värden in i det närmaste klustret, oavsett tidigare klustertillhörighet. Sedan beräknas nya mittpunkter för klustren, här används medeltalet av klustrets värden. Detta repeteras ända tills klustren inte mera utbyter värden [3]. Illustration 3.2 visar ett exempel på hur diskretisering med klustermetoden går till.

Klustermetoden hanterar obalanserade mängder, till skillnad från medianmetoden. Dock är klustermetoden väldigt beroende av värdenas fördelning. Bäst resultat fås från mängder där värdena tillhör två naturliga grupper [3]. Problem uppstår bland annat då det finns tre eller flera naturliga grupper eller då det minsta eller största värdet är så långt från medeltalet att de två sökta klustren slås ihop till ett.

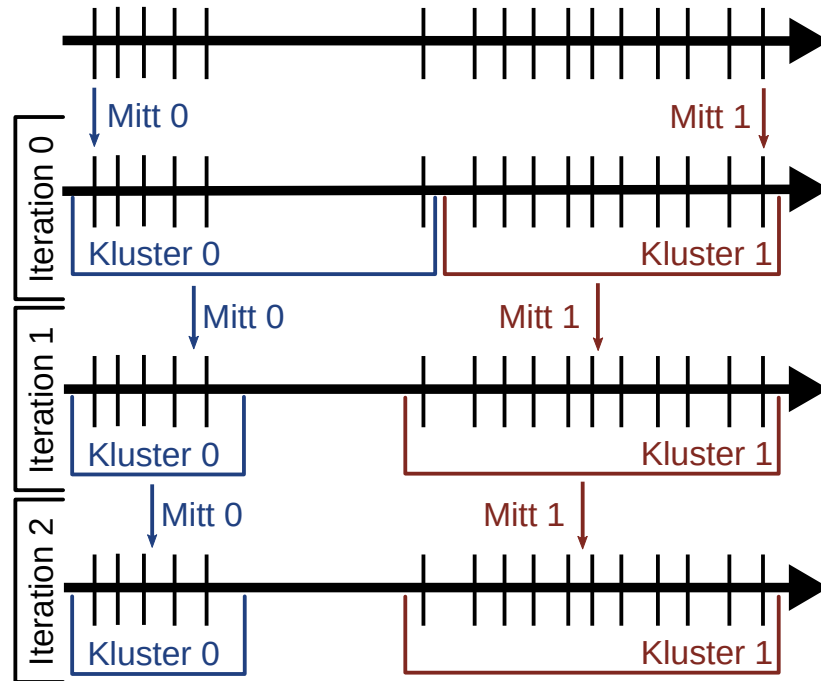


Illustration 3.2: Exempel på diskretisering med kluster

3.2 Signaturinlärning

Med alla värden omvandlade till booleska värden kan själva inlärningen börja. När inlärningen är klar förväntas resultatet vara signaturer i DNF, som beskrevs i avsnitt 2.2.1. Signaturerna är alltså mönster man kan jämföra egenskapernas värden mot för att se om ett utfall tillhör ett kluster.

Första steget i inlärningen är att välja vilka egenskaper som skall användas i signaturerna. Om egenskaperna som används för inlärning är för många jämfört med antalet sampel leder det till överanpassning av signaturerna. Detta innebär att resultatet att bli opålitligt för klassificering av nya sampel eftersom signaturerna kan (nästan) endast användas för att hitta de sampel som användes för att hitta signaturerna [13]. Detta går närmare igenom i nästa avsnitt.

Med ett mindre antal egenskaper kan signaturerna för samplen förkortas avsevärt. Detta är dock antagligen inte tillräckligt för att skapa behändiga

signaturer för klustren. För klustrens signaturer behövs ytterligare minimering, vilket beskrivs i avsnitten efter nästa.

3.2.1 Minimera antalet egenskaper

Om alla egenskaper tas i beaktande då signaturerna för klustren konstrueras kommer resultatet att bli en väldigt lång DNF. Signaturen ställer då krav på många irrelevanta egenskaper, vilket gör den alldeles för specifik för att användas för att klassificera nya sampel [13]. Dessutom är en lång signatur väldigt svårläst för en människa. Därför filtreras osignifikanta egenskaper bort innan signaturerna för klustren identifieras.

För att detta skall kunna göras krävs ytterligare information om hur viktiga de olika egenskaperna är. Till exempel en lista där egenskaperna är rangordnade enligt prioritet för det undersökta klustret. Med denna information söks det minsta antalet av de viktigaste egenskaperna som behövs för att skapa exklusiva kluster. Med exklusivitet menas att inget sampel i det undersökta klustret har samma signatur som ett sampel i ett annat kluster om signaturerna innehåller endast de valda egenskaperna. Dessa är de egenskaper som används för att hitta klustrets signatur i nästa avsnitt.

3.2.2 Signaturminimering

Med minimala egenskaperna kan ett klusters signatur skapas genom att disjunktivt kombinera signaturerna för klustrets sampel. Om klustret innehåller många sampel kommer signaturen vara väldigt lång. Längden gör att denna signatur blir otymplig att arbeta med. Kortare signaturer gör det både enklare att klassificera nya sampel, men framförallt blir det lättare för en människa att förstå och använda signaturerna. Därför tillämpas en algoritm för att förkorta signaturerna.

Av minimeringsprocessen önskas tre kvaliteter, exakthet, korthet och snabbhet. Med exakthet menas att den resulterande signaturen uttrycker exakt samma möjliga kombinationer som den ursprungliga signaturen, och inga fler. Hela idén med minimering är att hitta en kortare signatur, därför är

målet den kortaste möjliga signaturen. Dessutom föredras algoritmer som avslutas inom rimlig tid.

Med ett fåtal sampel och egenskaper kan Quine-McCluskey-algoritmen (QM) användas för att minimera signaturen. Med QM kan man garantera att den minimerade signaturen är den minsta möjliga. Dock är problemet med att hitta en minimal signatur NP-fullständigt. Det betyder att tiden det tar att utföra minimeringen växer exponentiellt mot antalet variabler och egenskaper. I de flesta fall behövs alltså en metod för att approximera resultatet [8].

En sådan algoritm är Espresso. Espresso är en heuristisk algoritm som genom iterationer konvergerar till ett lokalt minimum. Det finns alltså inget sätt att garantera att resultatet som fås från Espresso är det minsta möjliga. I praktiken leder dock Espresso oftast till bra resultat, som åstadkoms mycket snabbare än exakta metoder [8].

3.2.3 Espresso

Ursprungligen utvecklades Espresso för att optimera logikkretsar. Espresso utvecklades under början av 1980-talet och publicerades för första gången 1984. Sedan dess har flera modifieringar och implementationer presenterats [12]. I kapitel 2.2.1 beskrevs relationen mellan sanningstabeller och DNF. Denna relation är vad som möjliggör användningen Espresso, eftersom Essossos in- och utdata ofta är i form av sanningstabeller.

Espresso börjar med att bilda uttryck från samplens signaturer. Denna mängd av uttryck behandlas iterativt tills resultatet inte kan förbättras mera. Iterationen börjar med att uttrycken utvidgas, det vill säga begränsningar tas bort så att uttrycket täcker flera fall. Sedan tas uttryck, vars fall täcks av andra uttryck, bort. Som förberedelse inför nästa iteration förminskas de kvarvarande uttrycken. När flera iterationer inte förbättrar resultatet kan uttrycken drastiskt förändras i hopp om att ett lokalt minimum kan undvikas [12]. Nedan visas pseudokoden (3.1) för Essossos huvudsakliga struktur.

```

/*
  F är mängden av accepterade uttryck
  D är mängden av obestämda uttryck
  R är mängden av förkastade uttryck

  kostnad(F) prioriterar främst ett färre antal uttryck
              men tar också uttryckens längd i beaktande
*/

espresso(F,D,R)
{
  F ← EXPAND(F,R)      /* expanderar uttrycken */
  F ← IRREDUNDANT(F,D) /* tar bort överflödiga uttryck*/

  E ← ESSENTIAL(F,D)  /* hitta essentiella uttryck */
  F ← F - E           /* ta bort essentiella uttryck */
  D ← D + E

  do {
     $\varphi_2$  ← cost(F)

    /* repetera tills lösningen stabiliseras */
    do {
       $\varphi_1$  ← |F|

      F ← REDUCE(F,D) /* minimerar uttrycken */
      F ← EXPAND(F,R)
      F ← IRREDUNDANT(F,D)

    } while (|F| <  $\varphi_1$ )

    /* ändra lösningen för att undvika lokala minimum */
    G ← LAST_GASP(F,D,R)

  } while (kostnad(F) <  $\varphi_2$ )

  F ← F + E      /* lägg tillbaka essentiella uttryck*/
  D ← D - E

  return F
}

```

Pseudokod 3.1: Espressoalgoritmen huvudprocedur i pseudokod

I pseudokoden (3.1) tar Espresso emot tre mängder av fall, jakande, nekande och fall där resultatet inte spelar någon roll. Beroende på

implementationen behövs inte alla anges, då man genom komplement får den tredje mängden från två valfria mängder [12]. I inlärningsmetoden som undersöks i den här avhandlingen är varje kluster sin egen jakande mängd och den oberoende mängden tom. Detta gör att inlärningen inte gör antaganden om okända fall utan resultatet speglar inlärningsmaterialet exakt. Ett alternativ skulle vara att låta den oberoende mängden vara okänd och den nekande mängden innehålla alla sampel som finns i andra kluster än det vars signatur minimeras. Resultatet innehåller då kortare och läsligare uttryck. Dock kan okända fall klassificeras fel, och till och med höra till flera kluster.

4 Data

Metoden är dataagnostisk och fungerar på data från många olika källor och forskningsområden. Naturligtvis krävs dock att materialet har den form, eller kan transformeras till den form, som används i algoritmen. Exempel på denna förberedelse finns i avsnitt 3.1, där två metoder för att omvandla kontinuerliga värden till booleska presenteras. Dessutom behöver egenskaperna kunna rangordnas för att skapa relevanta signaturer, vilket förklaras närmare i avsnitt 3.2.1.

I denna avhandling kommer jag att använda öppet tillgänglig genuttrycksdata för att testa inlärningsmetoden. Genuttrycksdatan kommer från undersökningar om olika sjukdomar. I detta kapitel förklaras vad som avses med genuttryck och i slutet beskrivs dessutom källorna i närmare detalj.

4.1 Gener

Gener är sparade i DNA-molekyler i alla levande celler och definierar cellens och organismens, i det här fallet människans, egenskaper och funktioner. DNA står för deoxiribonukleinsyra och är en makromolekyl som styr produktionen av proteiner. DNA har en ryggrad av bestående av fosfat- och sockergrupper. I denna ryggrad sitter adenin (A), tymin (T), guanin (G) och cytosin (C) som är de molekyler som representerar data i DNA. DNA:s struktur är en dubbelspiral med två DNA-makromolekyler som speglar varandra då adenin och tymin sitter ihop (AT, TA) och guanin och cytosin sitter ihop (GC, CG). Det är följderna av dessa molekyler (A, T, G, C, t.ex. TTT) som bestämmer genens funktion [9].

Dubbelspiralens spegling (A-T och C-G) gör att ena DNA-makromolekylen i spiralen fungerar som mall för den andra, och vice versa. Det här möjliggör förmågan att kopiera DNA för celledelning eller produktion av proteiner. Vid celledelning separeras dubbelspiralen och två nya DNA-dubbelspiraler byggs upp med de nu åtskilda DNA-makromolekylerna som mall.

4.2 Genuttryck

Med genuttryck menas hur aktivt gener används för att producera protein. Detta är en viktig forskningsgren då mängden av de olika proteinerna som produceras avgör cellens funktion [10]. Produktionsprocessen för proteiner börjar med att DNA avläses till ribonukleinsyra (RNA). Med DNA som mall byggs mRNA (budbärar-RNA) som överförs till ribosomer där mRNA:t används för att skapa proteiner [9]. Att mäta mängden av olika sorters mRNA ger alltså mycket information om celler och organismer [10].

Mängderna av mRNA som bildas är inte endast beroende av generna. Till generna binds under organismens livstid reglerproteiner som stimulerar eller hämmar mRNA-produktionen för olika gener. Skapandet av dessa reglerproteiner beror på ämnen i omgivningen. Exempelvis hormoner, gifter, näringsämnen och celltypen kan orsaka förändringar i genuttrycken [10].

En mätning av genuttryck ger relativa värden. Det betyder att genuttryck för olika gener inte kan jämföras. Dessutom kan jämförelser mellan olika mätningsmetoder leda till felaktiga resultat [1]. Detta har tagits i beaktande i kapitel 3.1 där diskretiseringen av data presenteras.

4.3 Material

Materialet kommer från tre olika undersökningar om olika sjukdomar. Dessa sjukdomar är bröstcancer, alzheimer samt huvud- och nackcancer. Datan är genuttrycksdata, som beskrevs i föregående avsnitt. Genuttrycken är tagna från patienter i olika sjukdomstillstånd, exempelvis frisk kontra insjuknad i alzheimer. [11]

Då inlärningsmetoden appliceras på genuttrycken fås signaturer, mönster, över aktiva och inaktiva gener. Dessa signaturer gäller för olika sjukdomstillstånd, som används som bas för klustren i datan. Genuttryckens värden används för att avgöra huruvida en gen är aktiv eller inaktiv. Tabell 4.3.1 visar hur genuttryckstermer relaterar till inlärningsmetodens termer i kapitel 3.

Inlärningsmetod	Genuttryck
Material / Data	Genuttrycksdata
Kluster	Sjukdomstillstånd
Sampel	Patient
Egenskap	Gen
Värde	Genuttryck
Booleska värden	Aktiv / inaktiv gen

Tabell 4.3.1: Relationen mellan termer i inlärningsmetoden och genuttrycksdatan

Materialet har hämtats från 'Gene Expression Omnibus' (GEO), som är en databas för öppet tillgänglig genuttrycksdata. Dessutom erbjuder GEO ett gemensamt filformat som gör det lätt att byta mellan olika datamängder [2]. I tabell 4.3.2 är de olika sjukdomarna och detaljer om deras genuttrycksdata uppställd. Dessa datamängder kommer användas i nästa kapitel då inlärningsmetoden används för att hitta signaturer.

Sjukdom	Kluster	Sampel	Källa
Bröstcancer	4	47	[11]
Alzheimer	2	161	[7]
Huvud- och nackcancer	8	138	[14]

Tabell 4.3.2: Information om materialet

5 Inläring

För att utvärdera inlärningsmetoden har metodiken från kapitel 3 implementerats och använts på materialet från kapitel 4. Resultatet från inläringen presenteras i avsnitt 5.2. Detta följs upp av en utvärdering om inlärningsmetodens precision där en del av samplen har avskiljts till en testdatamängd. Först inleder dock avsnitt 5.1 med några detaljer om själva implementationen.

5.1 Implementation

Implementationen följer den metodik som presenterades i kapitel 3. Valet av programmeringsspråk är relativt trivialt. Självt valde jag java. Då stora datamängder behandlas skall rekursioner som är beroende av datamängden undvikas då det lätt leder till att stacken flödar över. Med språk med dynamisk minnesanvändning, såsom java, bör man dessutom undvika att överbelasta skräpsamlaren med många temporära objekt.

Ytterligare optimering fås om beräkningar parallelliseras genom användandet av flera trådar. Detta är möjligt då beräkningar för enskilda element, egenskaper, sampel eller kluster, inte beroende av resultat från beräkningar för andra element. En enkel lösning, som också skalar väldigt bra, är att använda sig av arbetartrådar. Arbetartrådarna tar emot beräkningsarbeten från en huvudtråd som har hand om programmets helhetsstruktur.

5.2 Resultat

För alla källor har följande information getts till inlärningsmetoden, datafil, kluster och de gener som är viktiga för den undersökta sjukdomen. Om dessa gener inte räcker till för att separera klustren (se avsnitt 3.2.1) används slumpmässigt valda gener. Tabell 5.2.1 visar vilka gener som är viktiga för de olika sjukdomarna.

Bröstcancer [11]	GATA3 MET HGF BRCA1 LDHB Erbb2 TP53 PRB1 CDKN2A EGFR AKT1
Alzheimer [7]	APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2
Huvud- och nackcancer [14]	CKAP5 KPNB1 RAN TPX2 KIF11

Tabell 5.2.1: Essentiella gener för de olika sjukdomarna

Då de slumpmässigt valda generna varierar mellan körningarna av programmet kommer resultatet som presenteras här vara det bästa från sex körningar. Med bästa avses det resultat med lägsta antalet gener som behövs för signaturerna. Tabellerna 5.2.2, 5.2.3 och 5.2.4 visar en översikt av inläringens resultat. Själva inlärd signaturerna hittas i bilagorna [1](#), [2](#) och [3](#) eftersom de ibland är väldigt långa.

Kluster	Sampel	Gener	Konjunktioner
Normal	7	10	3
Non-basal-like cancer	20	12	15
BRCA1-associated cancer	2	10	2
Basal-like cancer	18	12	16

Tabell 5.2.2: Resultatet från inläringen av bröstcancersignaturer, se [bilaga 1](#) för klustrens signaturer

Kluster	Sampel	Gener	Konjunktioner
Alzheimers sjukdom	87	25	72
Frisk	74	25	59

Tabell 5.2.3: Resultatet från inläringen av alzheimersignaturer, se [bilaga 2](#) för klustrens signaturer

Kluster	Sampel	Gener	Konjunktioner
Cancer fas 0	2	23	2
Cancer fas I	8	19	7
Cancer fas II	14	23	14
Cancer fas III	28	30	28
Cancer fas IVa	77	30	70
Cancer fas IVb	6	26	6
Cancer fas IVc	1	16	1
Cancer fas okänd	2	15	2

Tabell 5.2.4: Resultatet från inläringen av huvud- och nackcancersignaturer, se [bilaga 3](#) för klustrens signaturer

Signaturerna för bröstcancerklustren baseras på tolv gener. Eftersom elva av dessa gener är essentiella så är resultatet ungefär så bra som möjligt. För de andra två sjukdomarna, Alzheimer samt huvud- och nackcancer, är förhållandet mellan använda gener och essentiella gener mycket mera ojämnt. Detta gör att de hittade signaturerna blir väldigt långa då de innehåller många gener som inte påverkar sjukdomen ifråga. Dock visar resultatet alzheimerdatan tecken på att kunna förbättras avsevärt om flera viktiga gener införs. Samma kan inte sägas om resultatet från tabell 5.2.4, eftersom där är klustrena alltför lika varandra.

5.3 Test

I föregående avsnitt presenterades resultatet från att använda inlärningsmetoden på genuttrycksdatan, men för att bättre kunna validera resultatet så måste inläringen testas. Ett vanligt sätt att testa övervakad maskininläring är att dela upp datamängden i två mängder, en som används för inläring och en som används för att testa resultatet [15]. Detta sätt har tillämpats även här. För signaturerna i föregående avsnitt (5.2) användes fullständiga kluster utan en testdatamängd.

Inlärningsmetodens precision är förmågan att korrekt klassificera okända fall. För att utreda precisionen har testandet gjorts flera gånger för att

åstadkomma ett mera heltäckande resultat. Varje gång har en ny, slumpmässigt vald, testdatamängd använts. Alla kluster har undersökts enskilt, det vill säga endast ett kluster åt gången har haft en delmängd separerad från klustret för testning. Samplen i denna delmängd kan efter inlärning klassificeras för att se om de tilldelas rätt, fel eller inget kluster alls.

I följande tabeller, 5.3.1, 5.3.2 och 5.3.3, visar resultatet från då inlärningsmetoden har utvärderats med genuttrycksdatan. Programmet har körts sex gånger, vilket resulterar i att olika gener har använts. De viktiga generna har dock alltid inkluderats. För varje körning har varje kluster utvärderats tre gånger med 30% av samplen som testmängd. Det innebär också att kluster med färre än tre sampel inte har utvärderats.

Kluster	Teststorlek	Korreakta	Okända	Felaktiga
normal	2 av 7	30	6	0
Non-basal-like cancer	0 av 2	0	0	0
BRCA1-associated cancer	6 av 20	56	52	0
Basal-like cancer	5 av 18	0	90	0

Tabell 5.3.1: Resultat från utvärdering av inlärningsmetoden med bröstcancerdata [11]. Testmängden var 30% av klusterstorleken och testet genomfördes 18 gånger med slumpmässigt valda sampel i testmängden.

Kluster	Teststorlek	Korreakta	Okända	Felaktiga
Alzheimers sjukdom	26 av 87	60	408	0
Frisk	22 av 74	59	337	0

Tabell 5.3.2: Resultat från utvärdering av inlärningsmetoden med alzheimerdata [7]. Testmängden var 30% av klusterstorleken och testet genomfördes 18 gånger med slumpmässigt valda sampel i testmängden.

Kluster	Teststorlek	Korreakta	Okända	Felaktiga
Cancer fas 0	0 av 2	0	0	0
Cancer fas I	2 av 8	0	36	0
Cancer fas II	4 av 14	6	66	0
Cancer fas III	8 av 28	28	116	0
Cancer fas IVa	23 av 77	162	252	0
Cancer fas IVb	1 av 6	0	18	0
Cancer fas IVc	0 av 1	0	0	0
Cancer fas okänd	0 av 2	0	0	0

Tabell 5.3.3: Resultat från utvärdering av inlärningsmetoden med huvud- och nackcancerdata [14]. Testmängden var 30% av klusterstorleken och testet genomfördes 18 gånger med slumpmässigt valda sampel i testmängden.

Testresultaten liknar väldigt mycket detaljerna kring signaturinläringen i föregående avsnitt. Bröstcancerklustrens signaturer kunde någotsånär användas för att identifiera okända fall. Basalliknande cancer är dock undantaget, inget testsampel klassificerades rätt. Orsaken till detta syns redan i signaturen för klustret som endast kunde förkortas marginellt. Detta betyder en väldig överanpassning av signaturen så att den nästan endast kan klassificera de sampel som användes för inläringen och inga andra.

För de andra två sjukdomarna är det överanpassning som gäller allt igenom. Igen är det något som kan spåras i de inlärdas signaturerna vars disjunktioner består av nästan lika många konjunktioner som antalet sampel i klustren. Intressant att notera är dock att inga felklassificeringar skedde under hela testet.

Eftersom inlärningsmetoden väldigt exakt bygger signaturer från inlärningsmaterialet är en hög andel okända klassificeringar för okända fall förväntade. Det är denna exakthet som gör att felklassificeringar undviks, men som visats ovan kan korrekta klassificeringar åstadkommas om informationen inlärningsdatan räcker till.

6 Avslutning

Nu när maskininlärningsmetoden har beskrivits och använts återstår att reflektera över hur bra metoden fungerar. Först utvärderas hur bra metoden uppfyller de mål som sattes upp för den. Därefter beaktas metodens begränsningar och brister. Med denna information kan ett slutgiltigt omdöme ges i avsnitt 6.3.

6.1 Utvärdering

För att kunna avgöra hur bra en maskininlärningsmetod är måste målen med inläringen definieras och utvärderas. Med den här inlärningsmetod en önskas hittas signaturer som definierar kluster, är möjliga att förstå och hittas inom rimlig tid. De följande avsnitten går in på detalj hur inlärningsmetoden lyckas följa dessa krav.

6.1.1 Signaturkvalitet

Från testresultaten i kapitel 5.3 kan utläsas att inlärningsmetoden har svårt att klassificera okända fall, men felklassificeringar är ändå ovanliga. Detta förklaras med hjälp av metodiken från kapitel 3 där det framkommer att de inlärdas signaturerna exakt representerar inlärningsmaterialet. Inlärningsmetoden bör således inte användas i förutsäggande syfte, utan för att undersökningar där de exakta signaturerna är eftertraktade.

6.1.2 Läslighet

Kapitel 2.2 förklarade hur DNF används för att skapa läsliga signaturer som resultat. I praktiken hittas ganska långa signaturer. Dessa är inte lätta att läsa, men de är möjliga att förstå. Vilket skilljer från några av de alternativa metoder som presenterades i kapitel 2.1, vars interna information inte är strukturerad med tanke på mänskligt användande. Dessutom kan en boolesk logisk normalform lätt visualiseras på sätt, eller omvandlas till uttryck, som är lättare att ta in, vilket också förklarades i kapitel 2.2.

6.1.3 Prestanda

Jag har kört inlärningsmetoden på både en kraftfull stationär dator och en svagare bärbar dator för att få en överblick över hur lång tid inläringen kan tänkas ta. För tidtagningen användes samma material som introducerades i kapitel 4. Alla körningar tog endast några sekunder, med den bärbara datorns tider drygt dubbelt så långa som den stationära datorns. Om programmet körs med bara en tråd istället för flera kan man också räkna med dubbelt så lång tid.

För alla körningar gick största delen av tiden åt till att parse datafilerna. Själva inläringen klarades av på mindre än en sekund, även för den största datamängden på den bärbara datorn. Detta gör att jag lugnt kan konstatera att inlärningsmetoden har bra prestanda.

6.2 Begränsningar

Inlärningsmetodens krav och begränsningar berör främst datamängden som används för maskininläringen. Viktigast är att all data måste kunna omvandlas till booleska värden för signaturerna i disjunktiv normalform. Strukturellt behövs då något som motsvarar en tabell med egenskaper, gener, på ena axeln och sampel på andra.

För inläringen behövs också information om materialet i form av färdigt definierade kluster som ska undersökas. Dessutom är kunskap om viktiga egenskaper, gener, essentiell för att få vettiga resultat, vilket tydligt syns i kapitel 5.

6.3 Slutsatser

Detta är en väldigt specialiserad maskininlärningsmetod. Metoden lämpar sig inte för att klassificera nya fall eller för att undersöka okänt material. För dessa tillämpningar kan till exempel något av de alternativ som presenterades i kapitel 2.1 användas.

Inlärningsmetoden passar bättre för tillämpningar där materialet är relativt känt och vad som söks är resultat som en människa kan läsa och använda för vidare forskning eller kunskap. Ett alternativt scenario som passar metoden är då resultatet bör vara exakt och inte använda andra antaganden än vad som specifikt finns i materialet. Detta återspeglas i kapitel 5, där materialet som inlärningsmetoden användes på fungerade bra då tillräckligt många viktiga gener var kända på förhand.

Referenser

- [1] Brazma Alvis, Vilo Jaak, Gene expression data analysis. FEBS Letters 480 , 2000
- [2] Edgar Ron, Domrachev Michael och Lash Alex E., Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research , 2002
- [2] Freitas Alex A., Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer, 2002
- [3] Furey Terrence S., Cristianini Nello, Duffy Nigel, Bednarski David W., Schummer Michèl, och Haussler David, Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics , 2000
- [4] Kriesel David, A Brief Introduction to Neural Networks, http://www.dkriesel.com/en/science/neural_networks, 2007, Hämtad 2016
- [5] Lehman Eric, Leighton F Tom, Meyer Albert R, Mathematics for Computer Science. MIT CSAIL, 2015
- [6] Liang WS, Dunckley T, Beach TG, Grover A et al., Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Data tillgänglig från NCBI GEO databasen, Skapad 10-07-2006, Hämtad 24-03-2016, <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281>
- [7] McGeer P. C., Sanghavi J. V., Brayton R. K. och Sangiovanni-Vicentelli A. L., ESPRESSO-SIGNATURE: a new exact minimizer for logic functions. IEEE Transactions on Very Large Scale Integration (VLSI) Systems vol. 1, no. 4, 1993
- [8] Nordén Bengt, Nordlund Stefan, DNA, <http://www.ne.se/uppslagsverk/encyklopedi/lång/dna>, Hämtad 2016-02-24

- [9] O'Connor C, Adams J, Essentials of Cell Biology. NPG Education, 2014
- [10] Richardson AL, Wang ZC, De Nicolo A, Lu X et al., X chromosomal abnormalities in basal-like human breast cancer. Data tillgänglig från NCBI GEO databasen, Skapad 2006-02-09, Hämtad 2016-02-29, <http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2250>
- [11] Rudell Richard L., Multiple-Valued Logic Minimization for PLA Synthesis, 1986, EECS Department, University of California, Berkeley
- [12] Triantaphyllou Evangelos, Data Mining and Knowledge Discovery via Logic-Based Methods. Springer, 2010
- [13] Walter V, Yin X, Wilkerson MD, Cabanski CR et al., Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. Data tillgänglig från NCBI GEO databasen, Skapad 16-07-2012, Hämtad 04-04-2016, <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39368>
- [14] Wu C, Rosenfeld R, Clermont G, Using Data-Driven Rules to Predict Mortality in Severe Community Acquired Pneumonia. PLoS ONE 9(4), 2014

Bilagor

Bilaga 1: Signaturer för bröstcancerkluster

Normal

$(MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR) \vee$
 $(\neg GATA3 \wedge MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR) \vee$
 $(\neg GATA3 \wedge MET \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR)$

Non-basal-like cancer

$(\neg GATA3 \wedge \neg MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR \wedge \neg ESRRA) \vee$
 $(GATA3 \wedge MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A$
 $\wedge EGFR \wedge \neg AKT1 \wedge \neg ESRRA) \vee (GATA3 \wedge \neg MET \wedge \neg HGF \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge$
 $\neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR \wedge AKT1 \wedge ESRRA) \vee (\neg MET \wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB$
 $\wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGFR \wedge AKT1 \wedge ESRRA) \vee (\neg GATA3 \wedge \neg MET$
 $\wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR \wedge \neg AKT1 \wedge$
 $\neg ESRRA) \vee (GATA3 \wedge \neg MET \wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge$
 $\neg CDKN2A \wedge \neg EGFR \wedge AKT1 \wedge \neg ESRRA) \vee (GATA3 \wedge \neg MET \wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB$
 $\wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR \wedge AKT1 \wedge ESRRA) \vee (GATA3 \wedge MET$
 $\wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge EGFR \wedge \neg AKT1 \wedge E$
 $SRRA) \vee (GATA3 \wedge \neg MET \wedge \neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge$
 $EGFR \wedge AKT1 \wedge ESRRA) \vee (GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53$
 $\wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGFR \wedge AKT1 \wedge \neg ESRRA) \vee (\neg GATA3 \wedge \neg MET \wedge \neg HGF \wedge \neg BRCA1$
 $\wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge AKT1 \wedge ESRRA) \vee (GATA3 \wedge \neg MET$
 $\wedge \neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGFR \wedge \neg AKT1 \wedge$
 $ESRRA) \vee (GATA3 \wedge \neg MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge$
 $\neg CDKN2A \wedge \neg EGFR \wedge \neg AKT1 \wedge ESRRA) \vee (GATA3 \wedge MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge$
 $\neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGFR \wedge \neg AKT1) \vee (GATA3 \wedge \neg MET \wedge \neg HGF \wedge$
 $\neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGFR \wedge AKT1 \wedge \neg ESRRA)$

BRCA1-associated cancer

$(GATA3 \wedge \neg MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R) \vee (\neg GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A$
 $\wedge \neg EGFR)$

Basal-like cancer

$(\neg GATA3 \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg$
 $AKT1 \wedge \neg ESRRRA) \vee$
 $(\neg GATA3 \wedge MET \wedge HGF \wedge \neg BRCA1 \wedge LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1$
 $\wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg AKT1 \wedge ESRRRA) \vee (\neg GATA3 \wedge MET \wedge \neg HGF \wedge BRCA1 \wedge LDHB \wedge$
 $\neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGF$
 $R \wedge AKT1 \wedge ESRRRA) \vee (\neg GATA3 \wedge \neg MET \wedge$
 $\neg HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg AKT1)$ \vee
 $(\neg GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge ERBB2 \wedge TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge$
 $\neg AKT1 \wedge ESRRRA) \vee (\neg GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge ERBB2 \wedge \neg TP53 \wedge$
 $\neg PRB1 \wedge \neg CDKN2A \wedge \neg EGF$
 $R \wedge \neg AKT1 \wedge \neg ESRRRA) \vee (\neg GATA3 \wedge MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge$
 $\neg TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge AKT1 \wedge ESRRRA) \vee (\neg GATA3$
 $\wedge MET \wedge HGF \wedge \neg BRCA1 \wedge LDHB \wedge ERBB2 \wedge TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGF$
 $R \wedge \neg AKT1 \wedge$
 $ESRRRA) \vee (\neg GATA3 \wedge \neg MET \wedge HGF \wedge \neg BRCA1 \wedge \neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge$
 $EGFR \wedge \neg AKT1 \wedge \neg ESRRRA) \vee (\neg GATA3 \wedge \neg MET \wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge \neg ERBB2 \wedge$
 $\neg TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg AKT1 \wedge ESRRRA) \vee$
 $(\neg GATA3 \wedge MET \wedge HGF \wedge BRCA1$
 $\wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge PRB1 \wedge \neg CDKN2A \wedge \neg EGF$
 $R \wedge AKT1 \wedge \neg ESRRRA) \vee (\neg GATA3$
 $\wedge \neg MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg AKT1$
 $\wedge \neg ESRRRA) \vee (\neg GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge \neg ERBB2 \wedge TP53 \wedge \neg PRB1 \wedge$
 $\neg CDKN2A \wedge EGF$
 $R \wedge AKT1 \wedge ESRRRA) \vee (\neg GATA3 \wedge MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge$
 $\neg ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge EGF$
 $R \wedge \neg AKT1 \wedge \neg ESRRRA) \vee (\neg GATA3 \wedge \neg MET$
 $\wedge HGF \wedge \neg BRCA1 \wedge \neg LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge \neg CDKN2A \wedge \neg EGF$
 $R \wedge AKT1 \wedge ESRRRA) \vee$
 $(\neg GATA3 \wedge \neg MET \wedge \neg HGF \wedge \neg BRCA1 \wedge LDHB \wedge ERBB2 \wedge \neg TP53 \wedge \neg PRB1 \wedge$
 $\neg CDKN2A \wedge \neg EGF$
 $R \wedge AKT1 \wedge ESRRRA)$

Bilaga 2: Signaturer för Alzheimerkluster

Alzheimer

(APOE A2M LRP1 GGA1 PIN1 SORL1 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

(APOE A2M LRP1 GGA1 PIN1 SORL1 PRDX2 241454_at 235823_at NRAP 243251_at C2orf67 229934_at 239157_at 238389_s_at 211429_s_at 1569009_s_at 229939_at PPHLN1 217107_at DLGAP4 TTPA MEOX2 FBXL16 MUDENG) v

67^229934_at^239157_at^238389_s_at^211429_s_at^1569009_s_at^229939_at^PPHLN1^217107_at
^DLGAP4^TTPA^MEOX2^FBXL16^MUDENG) v
(APOE^A2M^LRP1^GGA1^PIN1^SORL1^PRDX2^241454_at^235823_at^NRAP^243251_at^C2orf
67^229934_at^239157_at^238389_s_at^211429_s_at^1569009_s_at^229939_at^PPHLN1^217107_at^DL
GAP4^TTPA^MEOX2^FBXL16^MUDENG) v
(APOE^A2M^LRP1^GGA1^SORL1^PRDX2^241454_at^235823_at^NRAP^243251_at^C2orf67^2
29934_at^238389_s_at^211429_s_at^1569009_s_at^229939_at^PPHLN1^217107_at^DLGAP4^TTPA
^MEOX2^FBXL16^MUDENG) v
(APOE^A2M^LRP1^GGA1^PIN1^SORL1^PRDX2^241454_at^235823_at^NRAP^243251_at^C2orf
67^229934_at^239157_at^238389_s_at^211429_s_at^1569009_s_at^229939_at^PPHLN1^217107_at
^DLGAP4^TTPA^MEOX2^FBXL16^MUDENG)

Bilaga 3: Huvud- och nackcancersignaturer

Cancer fas 0

(¬CKAP5∧¬KPNB1∧¬RAN∧TPX2∧KIF11∧A_23_P342688∧¬NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧¬VWF∧¬LTA∧¬AHI1∧A_23_P74914∧¬PHF20∧A_32_P179083∧¬MTHFR∧¬TNN) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧¬TPX2∧¬KIF11∧A_23_P342688∧¬NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧¬VWF∧¬LTA∧¬AHI1∧A_23_P74914∧¬PHF20∧A_32_P179083∧MTHFR∧TNN)

Cancer fas 1

(¬CKAP5∧¬KPNB1∧¬RAN∧¬TPX2∧¬KIF11∧A_23_P342688∧TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧VWF∧LTA∧¬AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧¬TPX2∧¬KIF11∧A_23_P342688∧NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧VWF∧LTA∧¬AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧TPX2∧¬KIF11∧A_23_P342688∧NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧¬VWF∧¬LTA∧¬AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧TPX2∧KIF11∧A_23_P342688∧¬NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧VWF∧¬LTA∧¬AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧¬TPX2∧¬KIF11∧A_23_P342688∧NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧VWF∧LTA∧AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧TPX2∧¬KIF11∧A_23_P342688∧¬NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧¬VWF∧¬LTA∧AHI1∧A_23_P74914) ∨

(¬CKAP5∧¬KPNB1∧¬RAN∧¬TPX2∧¬KIF11∧A_23_P342688∧¬NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧¬VWF∧LTA∧AHI1∧A_23_P74914)

Cancer fas 2

(¬CKAP5∧¬KPNB1∧¬RAN∧TPX2∧KIF11∧A_23_P342688∧NM_025057_1_1971∧¬TSPAN15∧¬ARCN1∧¬DHRS3∧¬TGFA∧¬TMEM154∧A_23_P323488∧¬RB1∧¬SH2D1A∧VWF∧¬LTA

S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_3
 2_P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS3Λ
 TGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_P
 179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
 3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_P
 179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
 3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_P1
 79083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
 3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_
 P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
 (CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
 S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAHI1ΛA_23_P74914ΛPHF20ΛA_32_P1

79083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_P
179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_P
179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHRS
3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_P
179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_
P179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √
(¬CKAP5ΛKPNB1ΛRANΛTPX2ΛKIF11ΛA_23_P342688ΛNM_025057_1_1971ΛTSPAN15ΛARCN1ΛDHR
S3ΛTGFAΛTMEM154ΛA_23_P323488ΛRB1ΛSH2D1AΛVWFΛLTAΛAH11ΛA_23_P74914ΛPHF20ΛA_32_P
179083ΛMTHFRΛTNNΛC9orf6ΛTEX15ΛSEPT6ΛMLF1ΛEAF1ΛTNRC6BΛSEC11A) √

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6^MLF1^EAF1^TNRC6B^SEC11A) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6^MLF1^EAF1^TNRC6B^SEC11A) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6^MLF1^EAF1^TNRC6B^SEC11A)

Cancer fas 4b

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF^LTA^AHI1^A_23_P74914^PHF20^A_32_P179083^MTHFR^TNN^C9orf6^TEX15^SEPT6)

Cancer fas 4c

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A^VWF)

Cancer fas okänd

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A) v

(CKAP5^KPNB1^RAN^TPX2^KIF11^A_23_P342688^NM_025057_1_1971^TSPAN15^ARCN1^DHRS3^TGFA^TMEM154^A_23_P323488^RB1^SH2D1A)