行政院國家科學委員會專題研究計畫 成果報告

癌症細胞生化路徑網路的交互作用(第 2 年)
研究成果報告(完整版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 96-2221-E-468-012-MY2
執 行 期 間 ：97 年 08 月 01 日至 98 年 07 月 31 日
執 行 單 位 ：亞洲大學生物科技學系

計 畫 主 持 人：張培均

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 98 年 11 月 07 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告
□ 期中進度報告

癌症細胞生化路徑網路的交互作用

計畫類別：■ 個別型計畫　　□ 整合型計畫

計畫編號：NSC 96-2221-E-468-012-MY2

執行期間：　96　年 8 月 1 日至　98　年 7 月 31 日

計畫主持人：張培均
共同主持人：
計畫參與人員：蕭雅莉、石貴中、李崇鴻、陳怡君、林書韻、張宜婷

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年■二年後可公開查詢

執行單位：亞洲大學生物資訊學系

中　華　民　國　九十八　年　十　月　三十一　日

中文摘要

關鍵詞：癌症，生化路徑網路，DNA 微陣列，基因次序結構，關連分析

細胞內的生化路徑網路是一體的，雖然可分為基因調控網路、訊號傳遞網路、蛋白質交互作用網路以及代謝網路等等的類別，每個類別又可分出具有特定特徵或功能的次網路，次網路之間是相連的，存在交互作用，甚至不同類別的生化路徑次網路也有關連存在，然而，針對癌症細胞中這種交互作用的分析文獻，目前仍未見到。

本計畫在研究癌細胞中生化路徑次網路之間的交互作用，首先以目前已知的生化路徑網路為基礎，重新整理並定義其次網路以及包含之基因，再利用不同細胞狀態具有不同的基因網路次序結構，以DNA 微陣列數據來篩檢不同癌症細胞的特徵基因，最後利用關連分析的方法，做生化路徑次網路與癌症細胞特徵基因的關連分析。

Keywords: cancer, pathway, DNA microarray, gene ordering structure, association analysis.

There are many types of pathways in the cell cooperate to maintain the cellular function exactly, such as gene regulation pathways, signaling pathways, protein interactions, and metabolic pathways. Each pathway network consists of many sub-networks that are characterized by special biochemical function or topology property. These sub-networks are associated by each other and existing interaction. Some interactions may occur between different types of pathways. However, the research literature of such interactions in cancer cell has never seen at present.

In this project, we studied the sub-network interactions in cancer cell. We intended to reorganize and annotate the sub-networks in pathways based on the public databases. Cancer related genes were screened according to the variations of gene ordering structures building by DNA microarray data in a tumor-versus-normal experiment. Association analysis was implement regarding cancer related genes and sub-network related genes to quantify the correlation.

研究目的

細胞的癌化是基因變異累積的結果，使得基因的表現或功能改變，進而細胞內整體基因間調控亦發生改變。癌細胞具有正常細胞所沒有的許多能力，例如；癌細胞的生長與細胞分裂是不受控制的，正常細胞在某些情況下會啟動自殺機制，癌細胞失去這種機制，癌細胞會刺激血管增生以獲得養分，癌細胞可以不斷做細胞分裂，癌細胞可以侵犯不同組織的其他器官，這些能力使癌症病人失去生命。

癌症相關基因是研究癌症的診斷、形成與治療的關鍵，近幾年來，基因微陣列技術被大量應用於癌症相關基因的研究，癌症相關基因的篩檢則是根據癌組織與正常組織之間，基因的表現量是否有顯著差異，以各種統計分析的方法來檢定，例如；倍數改變法（Fold-Change）(Schena *et al.*, 1995; DeRisi *et al.*, 1997; Chen *et al.*, 1997)、t-檢定法（t-test）、SAM(Tusher *et al.*, 2001)以及 Mann-Whitney-Wilcoxon Rank Sum 檢定(Chen *et al.*, 1997; Chambers *et al.*, 1999)等方法，然而，這些方法找出來的癌症相關基因，在數目上通常偏低，最近的相關文獻指出，癌症的產生是廣泛的基因變異的結果（Sjöblom *et al.*, 2006）。我們發展出利用基因表現的次序結構變異來做癌症分型的方法，並篩檢出癌症相關基因，結果證明其在癌症分類的靈敏度優於其他方法（Liu *et al.*, 2006），這表示次序結構變異的狀況，更能顯示細胞狀態的變異，而篩檢出來的癌症相關基因（次序變異基因對）應更有代表性。

研究背景

DNA 微陣列技術可以同時偵測數千甚至上萬個基因的表現,如此強而有力的工具被廣泛應用在癌症分類、癌症標記基因篩選、基因調控網路推論、功能基因體學等方面的研究(Winzeler *et al.*, 1999; Chen *et al.*, 2001; Yu *et al.,* 2004; Morley *et al.,* 2004; Chen *et al.,* 2005),其中,基因調控網路的推論與分析方法的發展,主要以時間序列的微陣列數據為分析對象,因為這樣的限制,這方面的研究目前以低等生物為主(Chen *et al.*, 2004; Liao *et al.*, 2005; Yugi *et al.*; 2005; Antonov *et al.*, 2006),這些生物在實驗操作上較容易。在癌症方面,癌症相關調控網路的推論,則大多必須結合其他類型的實驗分析方法,針對局部性少數幾個基因或蛋白質間,就其可能的關係做推論(Katoh *et al.*, 2005; Bommert *et al.*,2006; Ryan *et al.*, 2006),以系統性的方式來估計與癌症細胞相關的生化路徑網路,主要以統計上的關連性分析為主(Bussemaker *et al.*, 2001; Haverty *et al.*, 2004; Mlecnik *et al.*, 2005; Liu *et al.*, 2006),原因在於,目前的實驗技術並無法即時提供足夠的實驗數據,尤其是缺乏時間序列的數據,以作為癌症細胞相關生化路徑網路的定量分析。

在探討癌症相關的細胞內生化路徑網時,通常細胞內生化路徑網路是已知的,這些網路可以是基因調控網路、訊號傳遞網路、蛋白質交互作用網路以及代謝網路等。生物細胞內生化路徑網的公用資料庫十分豐富(Bader *et al.*, 2006),其類型分佈如下表:

Category	Number of databases
Protein–protein interactions	79
Metabolic pathways	43
Signaling pathways	41
Pathway diagrams	22
Transcription factors/gene regulatory networks	20
Protein–compound interactions	14
Genetic interaction networks	5
Protein sequence focused	12
Other	11

Unique total	196

以上這些公用資料庫，雖然都有詳細的註解，但是與人相關的部分共有 59 個資料庫，仍需再進一步整合。

細胞內生化路徑網路是由許多次網路(sub-network)組成，每個次網路具有特定的特徵或細胞功能，並且包含一群已知的特定基因，不同的癌症相關次網路可能不同，若以 S*m* 表示第 *m* 個次網路的基因群，其中包含 *k* 個基因，並以 G*n* 表示以微陣列篩檢癌症 *n* 所得的癌症相關基因群，假設此微陣列總共包含 *y* 個基因，而細胞內全部基因數為 *N*，令 *I*=S*m* ∩G*n* 的基因數，則對 S*m* 與 G*n* 這兩群基因的分佈做關連分析(Haverty *et al.*, 2004)，可計算出 p-value：

$$p-value=\sum_{i=I}^{k}\frac{C_i^y C_{k-i}^{N-y}}{C_k^N}$$

p-value 越小，表示癌症 *n* 與次網路 *m* 有較高的關連性，類似的關連性分析亦可用費氏精確檢定(Mlecnik *et al.*, 2005)。關於癌症相關基因群的篩檢方法，主要是比較癌症細胞與正常細胞的基因表現差異，做統計檢定，例如 t-檢定法（t-test）是最為常用的方法：

$$T_e=\frac{\left|\overline{x_1}-\overline{x_2}\right|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$S_p=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

其中 n_1 與 n_2 為癌症細胞與正常細胞的個別實驗次數，$\overline{x_1}$ 與 $\overline{x_2}$ 分別代表基因 *x* 在兩種細胞狀態的平均表現量，t-檢定通常要求實驗數據呈現常態分佈，因此，原始基因表現量數據須做對數轉換（$x\rightarrow\log(x)$），使接近常態分佈，若觀察值為 $T_{e,obs}$，則 p-vale 為 Prob($|T_e| > T_{e,obs}$)，一般定義 p-vale 小於 0.05 的基因為癌細胞相關基因。

我們在發展癌症分型方法的過程中，提出了以基因網路的次序結構來來定義細胞狀態並做癌症分型的方法(Liu *et al.*, 2006)。當兩個基因 *i* 與 *j* 在 DNA 微陣列上的信號強度具有顯著次序關係；即次序關係係數 r_{ij} 大於某一顯著性門檻值，則建立這兩個基因的連線：

$$\gamma_{ij} = 1 - \sum_{s=1}^{N} \frac{[x_{sj} - x_{si}]^+}{N(x_{\max} - x_{\min})}$$

$$[x_{sj} - x_{si}]^+ = \begin{cases} x_{sj} - x_{si} & x_{sj} > x_{si} \\ 0 & x_{sj} \le x_{si} \end{cases}$$

上式中 x_{si} 是實驗樣本s中基因i的表現信號強度；N 為樣本總數； $x_{\max}$ 與 $x_{\min}$ 則為實驗樣本中最大與最小基因表現信號強度。次序結構網路是一個有向圖(directed graph)，連線$i \rightarrow j$意味著基因i的信號強度比基因j來得低，我們定義i與j的連線為i流出而j流入，若i與j的連線具雙向性，我們定義為對等狀態。所以當某一個基因有很高的由內向外連線，表示這個基因有相對較低的表現信號強度；當某一個基因有很高的由外向內連線表示這個基因有相對較高的表現信號強度。因為我們不知道此次序關係係數r_{ij}的分佈，因此，門檻值的決定採用隨機取樣測試的過程來模擬樣本空間的分佈：(a) 隨機自各個實驗樣本中選取一對基因構成一組基因對，重複此過程5000次；(b) 對於這些隨機選取的各組基因對，分別計算次序關係係數 γ_p (p = 1, 2, … , 5000)；(c) 由上述可以獲得一個次序關係係數的分布，以這個次序關係係數分布的顯著水平1% (P < 0.01)做為建構次序結構網路的門檻值。

利用次序結構來評估細胞狀態的效果十分顯著，這可以從用於癌症分型的高精確度來證明(Liu *et al.*, 2006)，若比較癌症細胞與正常細胞的基因次序結構差異，可以篩檢出造成這些差異的相關基因，這些基因必定與癌症相關，這些基因的分佈是廣泛的，比由t-檢定等方法(Schena *et al.*, 1995; DeRisi *et al.*, 1997; Chen *et al.*, 1997; Chambers *et al.*, 1999; Tusher *et al.*, 2001)所篩檢出來的癌症相關基因，更有代表性。

細胞內的生化路徑網路是一體的，雖然，可分為基因調控網路、訊號傳遞網路、蛋白質交互作用網路以及代謝網路等等的類別，每個類別又可分出具有特定特徵或功能的次網路，然而，次網路之間是相連的，存在交互作用，甚至不同類別的生化路徑網路也有關連存在，針對癌症細胞中這種交互作用或關連的分析文獻，目前仍未見到。

我們將以目前已知的生化路徑網路為基礎，重新整理並定義其次網路以及包含之基因，再利用不同細胞狀態具有不同的基因網路次序結構，來篩選不同癌症細胞的特徵基因，最後利用現有的關連分析方法，做生化路徑次網路與癌症細胞特徵基因的關連分析，並發展分析方法來探討癌症相關生化路徑次網路間的彼此影響，以及生化路徑次網

路內節點之間，與癌症相關的可能互動關係。

研究方法

一、細胞生化路徑網路的整合

目前與人類有關的的生化路徑網路資料庫，可分為七大類，涵蓋蛋白質交互作用、代謝反應網路、信號傳遞網路、蛋白質-化合物交互作用網路以及基因調控網路，其資料庫名稱如下列表：

A. Protein-Protein Interactions

BIND - Biomolecular Interaction Network Database
BioGRID - General Repository for Interaction Datasets
CA1Neuron - Pathways of the hippocampal CA1 neuron Details
DIP - Database of Interacting Proteins
DopaNet - DopaNet
GPCR-PD - G protein-coupled receptors protein database
HiMAP - Human Interactome Map
HPID - Human Protein Interaction Database
HumanPSD - Human Proteome Survey Database
KinaseDB - Kinase Pathway Database
MINT - Molecular Interactions Database
OPHID - The Online Predicted Human Interaction Database
PhosphoSite - Cell Signaling Technology's PhosphoSite Database
PINdb - Proteins Interacting in the Nucleus database
POINT - Prediction of Interactome
PPID - Protein-Protein Interaction Database
ProChart - ProChart database of signal transduction pathway information
ProNet - Protein-protein Interaction Database
PubGene - PubGene
Ulysses - Projection of Protein Networks across Species
HiMAP - Human Interactome Map
Cancer Cell Map - The Cancer Cell Map

B. Metabolic Pathways

GenMAPP - Gene MicroArray Pathway Profiler
GOLD.db - Genomics of Lipid-associated Disorders
MetaCore - MetaCore pathway database
PathArt - Pathway Articulator
Reactome - Reactome KnowledgeBase

Reactome - Reactome KnowledgeBase
GenMAPP - Gene MicroArray Pathway Profiler
PathArt - Pathway Articulator
GOLD.db - Genomics of Lipid-associated Disorders
MetaCore - MetaCore pathway database

C. Signaling Pathways

CA1Neuron - Pathways of the hippocampal CA1 neuron
Cancer Cell Map - The Cancer Cell Map
GenMAPP - Gene MicroArray Pathway Profiler
GOLD.db - Genomics of Lipid-associated Disorders
Hedgehog - Hedgehog Signaling Pathway Database
INOH - Integrating Network Objects with Hierarchies
MetaCore - MetaCore pathway database
PANTHER - PANTHER
PathArt - Pathway Articulator
Pathways Knowledge Base - Ingenuity Pathways Knowledge Base
PhosphoSite - Cell Signaling Technology's PhosphoSite Database
PID - CMAP Pathway Interaction Database
Reactome - Reactome KnowledgeBase
TRMP - Therapeutically Relevant Multiple Pathways Database

D. Pathway Diagrams

BioCarta - BioCarta Pathway Diagrams
Hedgehog - Hedgehog Signaling Pathway Database
INOH - Integrating Network Objects with Hierarchies
PID - CMAP Pathway Interaction Database
TRMP - Therapeutically Relevant Multiple Pathways Database

E. Transcription Factors / Gene Regulatory Networks

cisRED - Cis-regulatory element database
ECRbase - Evolutionary conserved region database
Hedgehog - Hedgehog Signaling Pathway Database
HemoPDB - Hematopoiesis Promoter Database
MAPPER - MAPPER
microrna.org - Microrna.org target database
TRED - Transcriptional Regulatory Element Database

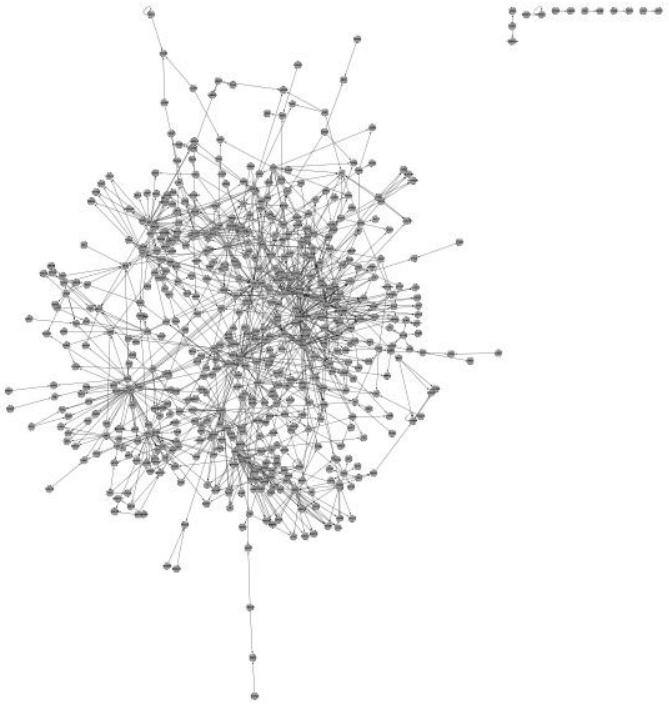
F. Protein-Compound Interactions

CTD - Comparative Toxicogenomics
ORDB - Olfactory Receptor Database

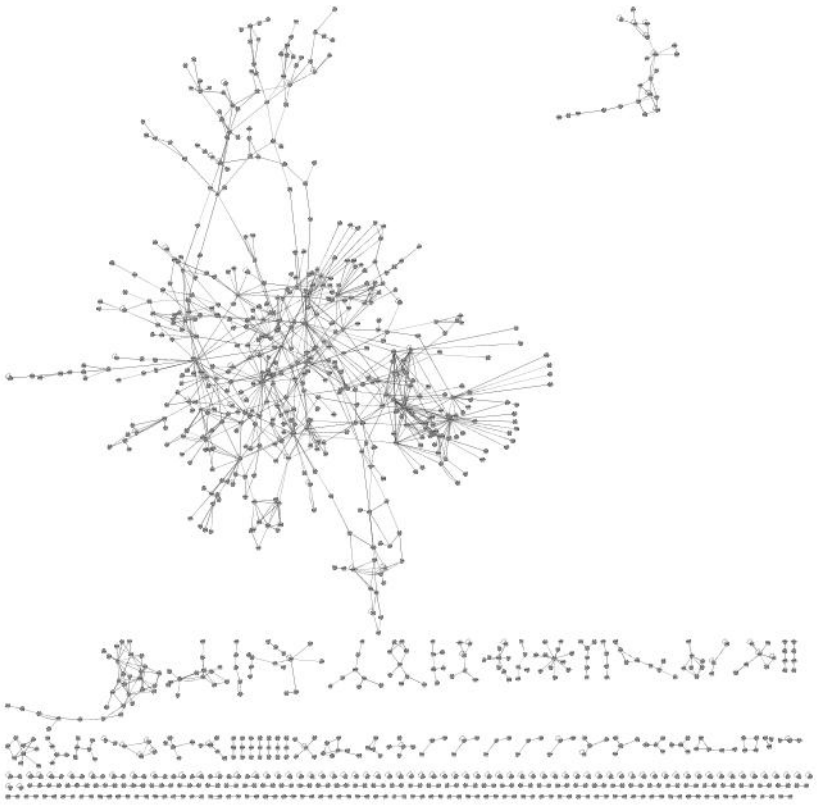
G. Genetic Interaction Networks

BIND - Biomolecular Interaction Network Database
BioGRID - General Repository for Interaction Datasets

以上這些資料庫有些同時含有不同類別之生化路徑網路資料，有些必須付費才能使用，我們篩選了較完整的資料庫並整合成人類的基因調空網路以及蛋白質交互作用網路，以人工單筆查詢的方式，逐一下載整理，原始資料中註解有誤或不一致的基因或蛋白質均予刪除，最後得到如下圖所示的生物網路。



人類基因調控網路



人類蛋白質交互作用網路

以上的網路是到目前為止，我們所整合出來最完整的生化路徑網路，若以三點以上成一子群，則其中包含 56 個基因調控子網路以及 135 個蛋白質交互作用子網路，由於人類相關的資料仍不夠豐富，真實的網路應該更為複雜。至於代謝反應的生物網路，則以 KEGG 的資料庫為主，並針對其既有之分類為標準，作為後續癌症相關代謝網路分析的依據。

二、微陣列數據資料庫來源

我們使用 Oncomine database 作為微陣列基因表現的數據來源(Rhodes *et al.*, 2004)，這是專門提供作為癌症基因體研究的資料庫，其中包含 18,740 個 DNA 微陣列實驗，約 58,000,000 筆以上的基因表現數據。Oncomine database 蒐集的這些微陣列資料並不提供直接下載，但是有提供其來源可以作為蒐集資料的線索，其中包含 cDNA 微陣列數據與寡核甘酸微陣列數據。相關的人類癌症微陣列實驗數據來源如下表：

Database	Web site	Reference
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	Brazma *et al.*, 2005
GeneNote	http://genecards.weizmann.ac.il/genenote/	Shmueli *et al.*, 2003

GEO	http://www.ncbi.nlm.nih.gov/geo/	Edgar *et al.*, 2002
HugeIndex	http://zlab.bu.edu/HugeSearch	Haverty *et al.*, 2002
ITTACA	http://bioinfo.curie.fr/ittaca	Elfilali *et al.*, 2006
LOLA	http://www.lola.gwu.edu/	
PEPR	http://pepr.cnmcresearch.org	Chen *et al.*, 2004
RefExA	http://www.lsbm.org/site_e/database/index.html	
SOURCE	http://source.stanford.edu	Diehn *et al.*, 2003
SMD	http://genome-www.stanford.edu/microarray	Ball *et al.*, 2005

蒐集所得的微陣列數據均加以註解，註解內容包含基因的 GenBank accession numbers、RefSeq ID、UniGene ID、Gene Ontology ID、所屬的生化路徑次網路以及基因功能描述等，並將每一個 DNA 微陣列實驗的基因表現對應到生化路徑網路上。

根據 Oncomine database 這個資料庫提供的資訊，我們蒐集整理了 126 個微陣列實驗資料集，同一組實驗資料集往往針對多種類型的癌症，總共涵蓋 18 種類型、143 種亞型的癌症，如下列表：

#	Type	Subtype	Platform*	Reference*
1	Bladder Cancer	Infiltrating Bladder Urothelial Carcinoma	Human Genome U133A Array	Dyrskjot *et al.*
		Stage 0is Bladder Urothelial Carcinoma	Human Genome U133A Array	Dyrskjot *et al.*
		Superficial Bladder Cancer	Human Genome U133A Array	Dyrskjot *et al.*
2	Brain and CNS Cancer	Anaplastic Astrocytoma	Human Genome U133 Plus 2.0 Array	Sun *et al.*
		Anaplastic Oligoastrocytoma	Human Genome U133 Plus 2.0 Array	French *et al.*
		Anaplastic Oligodendroglioma	Human Genome U133 Plus 2.0 Array	French *et al.*
		Astrocytoma	Human Genome U95A-Av2 Array	Shai *et al.*
		Atypical Teratoid/Rhabdoid Tumor	HumanGeneFL Array	Pomeroy *et al.*
		Classic Medulloblastoma	HumanGeneFL Array	Pomeroy *et al.*
		Desmoplastic Medulloblastoma	HumanGeneFL Array	Pomeroy *et al.*
		Diffuse Astrocytoma	Human Genome U133 Plus 2.0 Array	Sun *et al.*
		Glioblastoma	Human Genome U133 Plus 2.0 Array	Lee *et al.*
		Meningioma	Human Cancer Biology Array	Watson *et al.*
		Oligoastrocytoma	Liang	Liang *et al.*
		Oligodendroglioma	Human Genome U133 Plus 2.0 Array	Sun *et al.*
		Pilocytic Astrocytoma	Human Genome U95A-Av2 Array	Gutmann *et al.*
		Primary Glioblastoma	Affymetrix GeneChip 100K SNP	Beroukhim *et al.*
		Secondary Glioblastoma	Affymetrix GeneChip 100K SNP	Beroukhim *et al.*
3	Breast Cancer	Ductal Breast Carcinoma in Situ	Radvanyi	Radvanyi *et al.*

		Ductal Breast Carcinoma	Human Genome U133 Plus 2.0 Array	Richardson *et al.*
		Fibroadenoma	Sorlie	Sorlie *et al.*
		Invasive Breast Carcinoma	Agilent Human Genome 44K	Finak *et al.*
		Invasive Lobular Breast Carcinoma	Human Genome U133 Plus 2.0 Array	Turashvili *et al.*
		Invasive Mixed Breast Carcinoma	Radvanyi	Radvanyi *et al.*
		Lobular Breast Carcinoma	Zhao	Zhao *et al.*
4	Cervical Cancer		Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
5	Gastrointestinal Cancer	Barrett's Esophagus	Hao Esophagus	Hao *et al.*
		Cecum Adenocarcinoma	Human Genome U133 Plus 2.0 Array	Kaiser *et al.*
		Colon Adenocarcinoma	GPL4811	Ki *et al.*
		Colon Adenoma	Human Genome U133 Plus 2.0 Array	Sabates-Bellver *et al.*
		Colon Carcinoma	Zou	Zou *et al.*
		Colon Mucinous Adenocarcinoma	Human Genome U133 Plus 2.0 Array	Kaiser *et al.*
		Colorectal Adenoma	GPL3408	Gaspar *et al.*
		Colorectal Carcinoma	Graudens	Graudens *et al.*
		Diffuse Gastric Adenocarcinoma	Chen	Chen *et al.*
		Esophageal Adenocarcinoma	Hao	Hao *et al.*
		Gastric Intestinal Type Adenocarcinoma	Chen	Chen *et al.*
		Gastric Mixed Adenocarcinoma	Chen	Chen *et al.*
		Rectal Adenocarcinoma	Human Genome U133 Plus 2.0 Array	Kaiser *et al.*
		Rectal Adenoma	Human Genome U133 Plus 2.0 Array	Sabates-Bellver *et al.*
		Rectal Mucinous Adenocarcinoma	Human Genome U133 Plus 2.0 Array	Kaiser *et al.*
		Rectosigmoid Adenocarcinoma	Human Genome U133 Plus 2.0 Array	Kaiser *et al.*
6	Head and Neck Cancer	Floor of the Mouth Carcinoma	Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
		Head and Neck Squamous Cell Carcinoma	Human Genome U133A Array	Ginos *et al.*
		Hypopharyngeal Squamous Cell Carcinoma	Human Genome U133A Array	Schlingemann *et al.*
		Oral Cavity Carcinoma	Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
		Oral Cavity Squamous Cell Carcinoma	Human Genome U133A Array	Toruner *et al.*
		Oropharyngeal Carcinoma	Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
		Salivary Gland Adenoid Cystic Carcinoma	Human Genome U95A-Av2 Array	Frierson *et al.*
		Thyroid Gland Carcinoma	Human Genome U95A-Av2 Array	Huang *et al.*
		Tongue Carcinoma	Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
		Tongue Squamous Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Ye *et al.*
		Tonsillar Carcinoma	Human Genome U133 Plus 2.0 Array	Pyeon *et al.*
7	Kidney Cancer	Chromophobe Renal Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Yusenko *et al.*
		Clear Cell Renal Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Yusenko *et al.*
		Clear Cell Sarcoma of the Kidney	Human Genome U133A Array	Cutcliffe *et al.*
		Granular Renal Cell Carcinoma	Higgins	Higgins *et al.*
		Hereditary Clear Cell Renal Cell Carcinoma	Human Genome U133A Array	Beroukhim *et al.*

		Non-Hereditary Clear Cell Renal Cell Carcinoma	Human Genome U133A Array	Beroukhim *et al.*
		Papillary Renal Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Yusenko *et al.*
		Renal Wilms Tumor	Human Genome U133 Plus 2.0 Array	Yusenko *et al.*
8	Leukemia	Acute Adult T-Cell Leukemia/Lymphoma	Human Genome U133A Array	Choi *et al.*
		Acute Myeloid Leukemia	Human Genome U133A Array	Valk *et al.*
		B-Cell Acute Lymphoblastic Leukemia	Andersson	Andersson *et al.*
		Chronic Adult T-Cell Leukemia/Lymphoma	Human Genome U133A Array	Choi *et al.*
		Chronic Lymphocytic Leukemia	Human Genome U95A-Av2 Array	Haslinger *et al.*
		Hairy Cell Leukemia	Human Genome U95A-Av2 Array	Basso *et al.*
		T-Cell Acute Lymphoblastic Leukemia	Andersson	Andersson *et al.*
		T-Cell Prolymphocytic Leukemia	Human Genome U133A Array	Durig *et al.*
9	Liver Cancer	Cirrhosis	Human Genome U133 Plus 2.0 Array	Wurmbach *et al.*
		Focal Nodular Hyperplasia of the Liver	Chen	Chen *et al.*
		Hepatocellular Adenoma	Chen	Chen *et al.*
		Hepatocellular Carcinoma	Human Genome U133 Plus 2.0 Array	Wurmbach *et al.*
		Liver Cell Dysplasia	Human Genome U133 Plus 2.0 Array	Wurmbach *et al.*
10	Lung Cancer	Large Cell Lung Carcinoma	Garber	Garber *et al.*
		Lung Adenocarcinoma	Human Genome U133A Array	Su *et al.*
		Lung Carcinoid Tumor	Human Genome U95A-Av2 Array	Bhattacharjee *et al.*
		Small Cell Lung Carcinoma	Garber	Garber *et al.*
		Squamous Cell Lung Carcinoma	Garber	Garber *et al.*
11	Lymphoma	Activated B-Cell-Like Diffuse Large B-Cell Lymphoma	Alizadeh	Alizadeh *et al.*
		Burkitt's Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
		Centroblastic Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
		Cutaneous Follicular Lymphoma	Storz	Storz *et al.*
		Diffuse Large B-Cell Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
		Follicular Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
		Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma	Alizadeh	Alizadeh *et al.*
		Mantle Cell Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
		Marginal Zone B-Cell Lymphoma	Storz	Storz *et al.*
		Primary Effusion Lymphoma	Human Genome U95A-Av2 Array	Basso *et al.*
12	Melanoma	Benign Melanocytic Skin Nevus	Human Genome U133A Array	Talantov *et al.*
		Cutaneous Melanoma	Human Genome U133 Plus 2.0 Array	Riker *et al.*
		Non-Neoplastic Nevus	Haqq	Haqq *et al.*
13	Myeloma	Monoclonal Gammopathy of Undetermined Significance	Human Genome U133 Plus 2.0 Array	Zhan *et al.*
		Multiple Myeloma	HumanGeneFL Array	Zhan *et al.*

		Smoldering Myeloma	Human Genome U133 Plus 2.0 Array	Zhan *et al.*
14	Ovarian Cancer	Ovarian Adenocarcinoma	HumanGeneFL Array	Welsh *et al.*
		Ovarian Clear Cell Adenocarcinoma	Human Genome U95A-Av2 Array/Human Genome U95B Array/Human Genome U95C Array/Human Genome U95D Array/Human Genome U95E Array	Lu *et al.*
		Ovarian Endometrioid Adenocarcinoma	Human Genome U95A-Av2 Array/Human Genome U95B Array/Human Genome U95C Array/Human Genome U95D Array/Human Genome U95E Array	Lu *et al.*
		Ovarian Mucinous Adenocarcinoma	Human Genome U95A-Av2 Array/Human Genome U95B Array/Human Genome U95C Array/Human Genome U95D Array/Human Genome U95E Array	Lu *et al.*
		Ovarian Serous Adenocarcinoma	Human Genome U95A-Av2 Array/Human Genome U95B Array/Human Genome U95C Array/Human Genome U95D Array/Human Genome U95E Array	Lu *et al.*
		Ovarian Serous Cystadenocarcinoma	Human Genome U133A Array	http://cancergenome.nih.gov/
15	Pancreatic Cancer	Pancreatic Adenocarcinoma	Iacobuzio-Donahue	Iacobuzio-Donahue *et al.*
		Pancreatic Carcinoma	Human Genome U133A Array	Segara *et al.*
		Pancreatic Ductal Adenocarcinoma	Human Genome U133A Array/Human Genome U133B Array	Ishikawa *et al.*
		Pancreatic Intraepithelial Neoplasia	Buchholz Pancreas	Buchholz *et al.*
		Pancreatitis	HumanGeneFL Array	Logsdon *et al.*
16	Prostate Cancer	Benign Prostatic Hyperplasia	Tomlins	Tomlins *et al.*
		Prostate Adenocarcinoma	Human Genome U133A Array/Human Genome U133B Array	Vanaja *et al.*
		Prostate Carcinoma	Human Genome U133 Plus 2.0 Array	Varambally *et al.*
		Prostatic Intraepithelial Neoplasia	Tomlins	Tomlins *et al.*
17	Sarcoma	Dedifferentiated Liposarcoma	Human Genome U133A Array	Detwiller *et al.*
		Fibrosarcoma	Human Genome U133A Array	Detwiller *et al.*
		Leiomyosarcoma	Human Genome U133A Array	Detwiller *et al.*
		Malignant Fibrous Histiocytoma	Human Genome U133A Array	Detwiller *et al.*
		Pleomorphic Liposarcoma	Human Genome U133A Array	Detwiller *et al.*
		Round Cell Liposarcoma	Human Genome U133A Array	Detwiller *et al.*
		Synovial Sarcoma	Human Genome U133A Array	Detwiller *et al.*
		Uterine Corpus Leiomyosarcoma	HumanGeneFL Array	Quade *et al.*
18	Other Cancer	Actinic (Solar) Keratosis	Human Genome U133A Array	Nindl *et al.*

		Adrenal Cortex Adenoma	Human Genome U95A-Av2 Array	Giordano *et al.*
		Adrenal Cortex Carcinoma	Human Genome U95A-Av2 Array	Giordano *et al.*
		Embryonal Carcinoma	Human Genome U133A Array/Human Genome U133B Array	Korkola *et al.*
		Endometrial Endometrioid Adenocarcinoma	HumanGeneFL Array	Mutter *et al.*
		Familial Parathyroid Hyperplasia	Human Genome U133A Array	Morrison *et al.*
		Malignant Glioma	HumanGeneFL Array	Pomeroy *et al.*
		Mixed Germ Cell Tumor	Human Genome U133A Array/Human Genome U133B Array	Korkola *et al.*
		Non-Familial Multiple Gland Neoplasia	Human Genome U133A Array	Morrison *et al.*
		Parathyroid Gland Adenoma	Human Genome U133A Array	Morrison *et al.*
		Parathyroid Hyperplasia	Human Genome U133A Array	Morrison *et al.*
		Pleural Malignant Mesothelioma	Human Genome U133A Array	Gordon *et al.*
		Primitive Neuroectodermal Tumor	HumanGeneFL Array	Pomeroy *et al.*
		Renal Oncocytoma	Human Genome U133 Plus 2.0 Array	Yusenko *et al.*
		Seminoma	Human Genome U133A Array/Human Genome U133B Array	Korkola *et al.*
		Skin Basal Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Riker *et al.*
		Skin Squamous Cell Carcinoma	Human Genome U133 Plus 2.0 Array	Riker *et al.*
		Teratoma	Human Genome U133A Array/Human Genome U133B Array	Korkola *et al.*
		Testicular Intratubular Germ Cell Neoplasia	Skotheim	Skotheim *et al.*
		Testicular Seminoma	Skotheim	Skotheim *et al.*
		Testicular Teratoma	Skotheim	Skotheim *et al.*
		Testicular Yolk Sac Tumor	Skotheim	Skotheim *et al.*
		Uterine Corpus Leiomyoma	HumanGeneFL Array	Quade *et al.*
		Vulvar Intraepithelial Neoplasia	Human Genome U133 Plus 2.0 Array	Santegoets *et al.*
		Yolk Sac Tumor	Human Genome U133A Array/Human Genome U133B Array	Korkola *et al.*

*上表僅列出各類型癌症其中一個微陣列實驗與其參考資料。

三、癌症相關基因的篩檢方法

我們以次序結構變異來篩檢癌症相關基因。當兩個基因 i 與 j 在 DNA 微陣列上的信號強度具有顯著次序關係；即次序關係係數 r_{ij} 大於某一顯著性門檻值，則建立這兩個基因的連線：

$$\gamma_{ij} = 1 - \sum_{s=1}^{N} \frac{[x_{sj} - x_{si}]^+}{N(x_{\max} - x_{\min})}$$

$$[x_{sj} - x_{si}]^+ = \begin{cases} x_{sj} - x_{si} & x_{sj} > x_{si} \\ 0 & x_{sj} \le x_{si} \end{cases}$$

上式中 x_{si} 是實驗樣本 s 中基因 i 的表現信號強度；N 為樣本總數；$x_{\max}$ 與 $x_{\min}$ 則為實驗樣本中最大與最小基因表現信號強度。次序結構網路是一有向圖(directed graph)，連線 $i \rightarrow j$ 意味著基因 i 的信號強度比基因 j 來得低，所以當某一個基因有很高的由內向外連線，表示這個基因有相對較低的表現信號強度；當某一個基因有很高的由外向內連線表示這個基因有相對較高的表現信號強度。因為我們不知道此次序關係係數 r_{ij} 的分佈，因此，門檻值的決定採用隨機取樣測試過程：(a) 隨機自各個實驗樣本中選取一對基因構成一組基因對，重複此過程 5000 次；(b) 對於這些隨機選取的各組基因對，分別計算次序關係係數 γ_p (p = 1, 2, … , 5000)；(c) 由上述可以獲得一個次序關係係數的分布，以這個次序關係係數分布的顯著水平 1% (P < 0.01)做為建構次序結構網路的門檻值。

根據以上的演算法分別建立癌症以及與之相對應的正常細胞的次序結構網路，比較兩個次序結構，其中有次序關係變異的連線與其兩端的基因，即為癌症相關基因。

我們所收集的微陣列資料集有些是同類型癌症但是實驗平台或環境不同，若相同基因對在不同平台的實驗有不一致的次序關係變異，則分別針對 $i \rightarrow j$ 與 $j \rightarrow i$ 在癌細胞與正常細胞的 r 比值，先轉換為 ln(r 比值)，再利用 meta-analysis (Whitehead *et al.*, 1991)來檢定其是否顯著，並確認其次序關係變異是否存在，只要 $i \rightarrow j$ 與 $j \rightarrow i$ 其中之一的 ln(r 比值)是顯著的，即認定其次序關係變異存在，其兩端的基因，即為癌症相關基因。

四、癌症相關生化路徑次網路的關連性分析

利用次序關係變異所篩檢出來的癌症相關基因，對應到整個生物網路（包含基因調空網路、蛋白質交互作用網路以及代謝網路），並延伸其連結至下一層，定義為癌症相關生化網路，計算其拓樸性質並探討其可能存在的次網路。探討的基本拓樸性質如下：

1. Degree：自由度；是網路上節點的連線數量。這裡的連線數量在次序結構網路中，指的是向外連線與向內連線的總和。
2. Betweenness：居間度；是度量某節點被其他任兩節點間，最短路徑通過之次數。

3. Closeness：緊密度；是度量某節點至其他節點最短路徑的平均值。

若以 S 表示某個次網路的基因群，其中包含 k 個基因，以 G 表示以次序結構變異所篩檢出來的癌症相關基因群，假設此微陣列總共包含 y 個基因，而細胞內全部基因數為 N，令 I=S∩G 的基因數，則對 S 與 G 這兩群基因的分佈做關連分析(Haverty *et al.*, 2004)，可計算出 p-value：

$$p_{val} = \sum_{i=I}^{k} \frac{C_i^y C_{k-i}^{N-y}}{C_k^N}$$

p-value 越小，表示此癌症 G 與次網路 S 有較高的關連性，計算出 p-value 後，再計算出 Q-value (false discovery rate) (Storey *et al.*, 2003)。

超幾何分佈(hypergeometric probability)被廣泛應用在評估兩個事件的相關度，我們亦應用在評估不同類型癌症的相關性上，以瞭解不同癌症間的關連性，以建立癌症網路。

結果與討論

一、各類型癌症之間的相關性網路

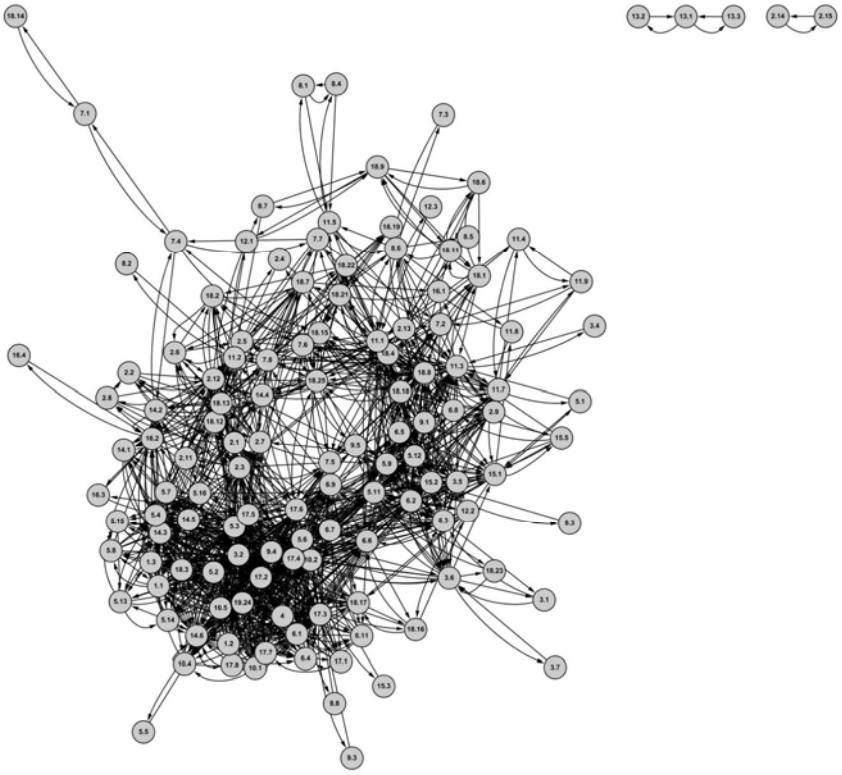
我們利用超幾何分佈(hypergeometric probability)來評估各類型癌症之間的相關程度，若兩種癌症的顯著相關基因數為 *Nr* 與 *Nq*，細胞內的基因總數為 *N*，且兩種類行癌症的特徵基因交集數為 *Nc*，則兩類型癌症相關程度的 p-value 計算如下式：

$$P_{val} = \sum_{i=Nc}^{Nq} \frac{C_{Nr}^{i} C_{N-Nr}^{Nq-i}}{C_{N}^{Nq}}$$

則相關性可定義如下：

$$Weight_{edge} = -\ln(P_{val})$$

若兩癌症類型間的 P_{val} 小於 0.01 既建立其連結關係，$Weight_{edge}$ 即為此連結的權重，如下圖：



結果顯示；同類型但是不同亞型的癌症的確較易聚在一起，但是有些不同類型的癌症反而比同類型癌症更有相關性，這表示不同類型的癌症有可能具有相同的生物網路模組或特徵基因，這表示其成癌的過程可能類似，在藥物治療上，可能也會有類似的反應，甚至可推論；相同的癌症藥物，也許可以治療不同類型的癌症，這個現象仍有待進一步加以探討。我們亦發現有兩小群的癌症沒有與其他癌症有顯著的關連，分別為黑色素瘤與神經膠質母細胞瘤。圖中的編號與各類型癌症的對照表如下表：

Type	Subtype	Index
Bladder Cancer	Infiltrating Bladder Urothelial Carcinoma	1.1
	Stage 0is Bladder Urothelial Carcinoma	1.2
	Superficial Bladder Cancer	1.3
Brain and CNS Cancer	Anaplastic Astrocytoma	2.1
	Anaplastic Oligoastrocytoma	2.2
	Anaplastic Oligodendroglioma	2.3
	Astrocytoma	2.4
	Atypical Teratoid/Rhabdoid Tumor	2.5
	Classic Medulloblastoma	2.6
	Desmoplastic Medulloblastoma	2.7
	Diffuse Astrocytoma	2.8
	Glioblastoma	2.9
	Meningioma	2.10
	Oligoastrocytoma	2.11
	Oligodendroglioma	2.12
	Pilocytic Astrocytoma	2.13
	Primary Glioblastoma	2.14
	Secondary Glioblastoma	2.15
Breast Cancer	Ductal Breast Carcinoma in Situ	3.1
	Ductal Breast Carcinoma	3.2
	Fibroadenoma	3.3
	Invasive Breast Carcinoma	3.4
	Invasive Lobular Breast Carcinoma	3.5
	Invasive Mixed Breast Carcinoma	3.6
	Lobular Breast Carcinoma	3.7
Cervical Cancer		4
Gastrointestinal Cancer	Barrett's Esophagus	5.1
	Cecum Adenocarcinoma	5.2
	Colon Adenocarcinoma	5.3
	Colon Adenoma	5.4
	Colon Carcinoma	5.5
	Colon Mucinous Adenocarcinoma	5.6
	Colorectal Adenoma	5.7
	Colorectal Carcinoma	5.8
	Diffuse Gastric Adenocarcinoma	5.9
	Esophageal Adenocarcinoma	5.10
	Gastric Intestinal Type Adenocarcinoma	5.11
	Gastric Mixed Adenocarcinoma	5.12

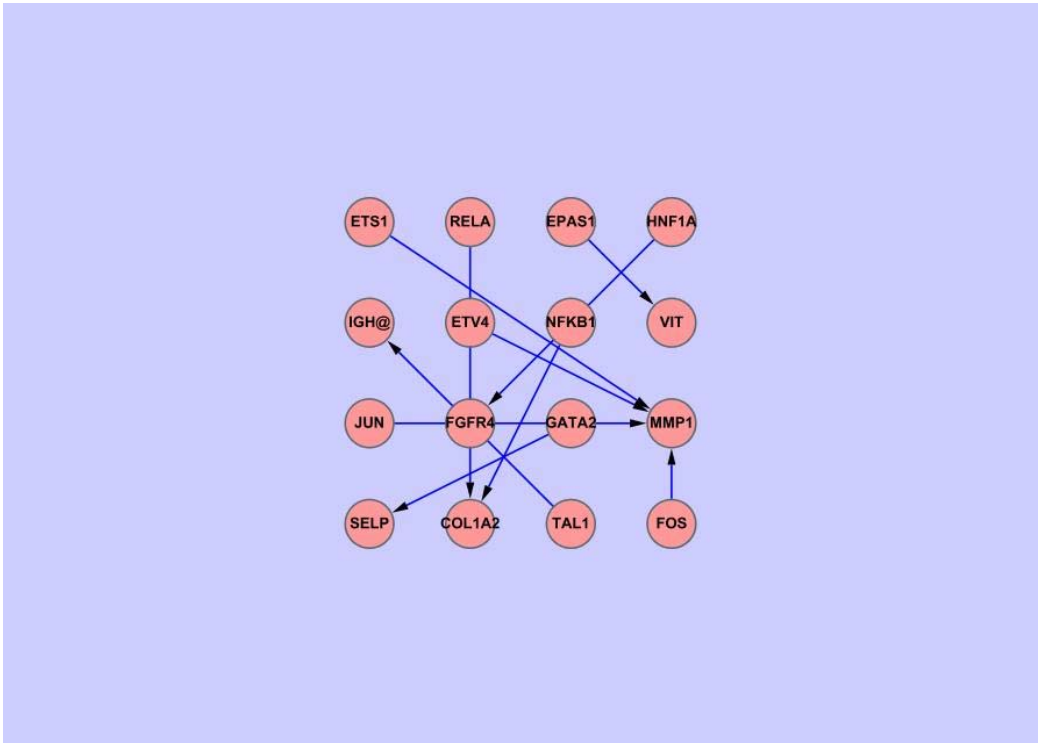
	Rectal Adenocarcinoma	5.13
	Rectal Adenoma	5.14
	Rectal Mucinous Adenocarcinoma	5.15
	Rectosigmoid Adenocarcinoma	5.16
Head and Neck Cancer	Floor of the Mouth Carcinoma	6.1
	Head and Neck Squamous Cell Carcinoma	6.2
	Hypopharyngeal Squamous Cell Carcinoma	6.3
	Oral Cavity Carcinoma	6.4
	Oral Cavity Squamous Cell Carcinoma	6.5
	Oropharyngeal Carcinoma	6.6
	Salivary Gland Adenoid Cystic Carcinoma	6.7
	Thyroid Gland Carcinoma	6.8
	Tongue Carcinoma	6.9
	Tongue Squamous Cell Carcinoma	6.10
	Tonsillar Carcinoma	6.11
Kidney Cancer	Chromophobe Renal Cell Carcinoma	7.1
	Clear Cell Renal Cell Carcinoma	7.2
	Clear Cell Sarcoma of the Kidney	7.3
	Granular Renal Cell Carcinoma	7.4
	Hereditary Clear Cell Renal Cell Carcinoma	7.5
	Non-Hereditary Clear Cell Renal Cell Carcinoma	7.6
	Papillary Renal Cell Carcinoma	7.7
	Renal Wilms Tumor	7.8
Leukemia	Acute Adult T-Cell Leukemia/Lymphoma	8.1
	Acute Myeloid Leukemia	8.2
	B-Cell Acute Lymphoblastic Leukemia	8.3
	Chronic Adult T-Cell Leukemia/Lymphoma	8.4
	Chronic Lymphocytic Leukemia	8.5
	Hairy Cell Leukemia	8.6
	T-Cell Acute Lymphoblastic Leukemia	8.7
	T-Cell Prolymphocytic Leukemia	8.8
Liver Cancer	Cirrhosis	9.1
	Focal Nodular Hyperplasia of the Liver	9.2
	Hepatocellular Adenoma	9.3
	Hepatocellular Carcinoma	9.4
	Liver Cell Dysplasia	9.5
Lung Cancer	Large Cell Lung Carcinoma	10.1
	Lung Adenocarcinoma	10.2
	Lung Carcinoid Tumor	10.3
	Small Cell Lung Carcinoma	10.4

	Squamous Cell Lung Carcinoma	10.5
Lymphoma	Activated B-Cell-Like Diffuse Large B-Cell Lymphoma	11.1
	Burkitt's Lymphoma	11.2
	Centroblastic Lymphoma	11.3
	Cutaneous Follicular Lymphoma	11.4
	Diffuse Large B-Cell Lymphoma	11.5
	Follicular Lymphoma	11.6
	Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma	11.7
	Mantle Cell Lymphoma	11.8
	Marginal Zone B-Cell Lymphoma	11.9
	Primary Effusion Lymphoma	11.10
Melanoma	Benign Melanocytic Skin Nevus	12.1
	Cutaneous Melanoma	12.2
	Non-Neoplastic Nevus	12.3
Myeloma	Monoclonal Gammopathy of Undetermined Significance	13.1
	Multiple Myeloma	13.2
	Smoldering Myeloma	13.3
Ovarian Cancer	Ovarian Adenocarcinoma	14.1
	Ovarian Clear Cell Adenocarcinoma	14.2
	Ovarian Endometrioid Adenocarcinoma	14.3
	Ovarian Mucinous Adenocarcinoma	14.4
	Ovarian Serous Adenocarcinoma	14.5
	Ovarian Serous Cystadenocarcinoma	14.6
Pancreatic Cancer	Pancreatic Adenocarcinoma	15.1
	Pancreatic Carcinoma	15.2
	Pancreatic Ductal Adenocarcinoma	15.3
	Pancreatic Intraepithelial Neoplasia	15.4
	Pancreatitis	15.5
Prostate Cancer	Benign Prostatic Hyperplasia	16.1
	Prostate Adenocarcinoma	16.2
	Prostate Carcinoma	16.3
	Prostatic Intraepithelial Neoplasia	16.4
Sarcoma	Dedifferentiated Liposarcoma	17.1
	Fibrosarcoma	17.2
	Leiomyosarcoma	17.3
	Malignant Fibrous Histiocytoma	17.4
	Pleomorphic Liposarcoma	17.5
	Round Cell Liposarcoma	17.6
	Synovial Sarcoma	17.7

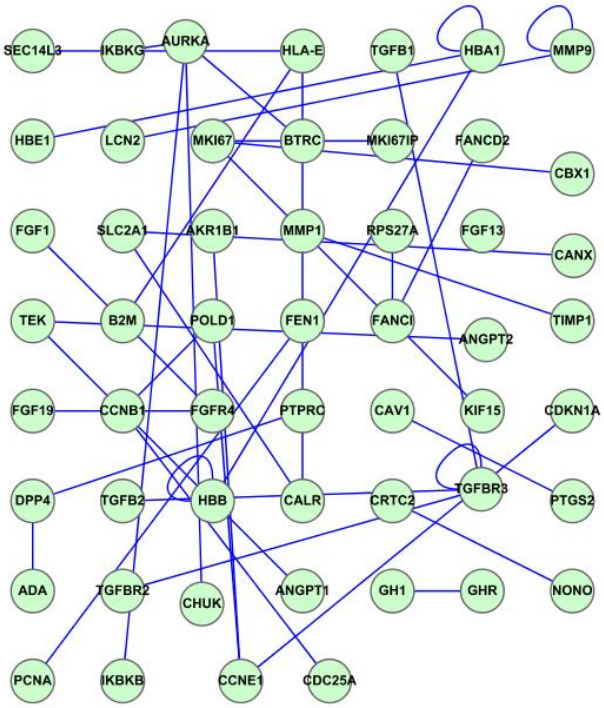
	Uterine Corpus Leiomyosarcoma	17.8
Other Cancer	Actinic (Solar) Keratosis	18.1
	Adrenal Cortex Adenoma	18.2
	Adrenal Cortex Carcinoma	18.3
	Embryonal Carcinoma	18.4
	Endometrial Endometrioid Adenocarcinoma	18.5
	Familial Parathyroid Hyperplasia	18.6
	Malignant Glioma	18.7
	Mixed Germ Cell Tumor	18.8
	Non-Familial Multiple Gland Neoplasia	18.9
	Parathyroid Gland Adenoma	18.10
	Parathyroid Hyperplasia	18.11
	Pleural Malignant Mesothelioma	18.12
	Primitive Neuroectodermal Tumor	18.13
	Renal Oncocytoma	18.14
	Seminoma	18.15
	Skin Basal Cell Carcinoma	18.16
	Skin Squamous Cell Carcinoma	18.17
	Teratoma	18.18
	Testicular Intratubular Germ Cell Neoplasia	18.19
	Testicular Seminoma	18.20
	Testicular Teratoma	18.21
	Testicular Yolk Sac Tumor	18.22
	Uterine Corpus Leiomyoma	18.23
	Vulvar Intraepithelial Neoplasia	18.24
	Yolk Sac Tumor	18.25

二、癌症特徵網路與次網路

利用次序結構差異篩選出來的癌症相關基因(p-value <0.01)，我們直接將這些基因以及其產物蛋白質對應到人類基因調控網路以及蛋白質交互作用網路，並延伸其連結的第一層，定義為癌症特徵基因調控網路與特徵蛋白質交互作用網路（附錄一）。例如下圖分別為肺腺癌(Lung Adenocarcinoma)的特徵基因調控網路與特徵蛋白質交互作用網路：



肺腺癌特徵基因調控網路

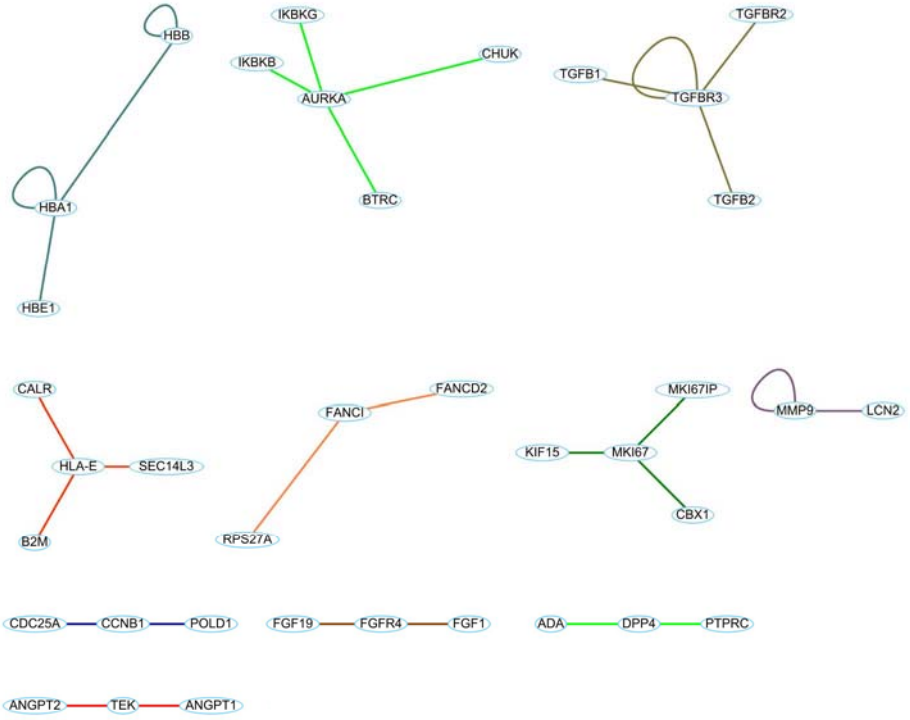


肺腺癌特徵蛋白質交互作用網路

在肺腺癌特徵基因調控網路中，有兩個明顯的集散點(hub)，分別為 COL1A2 與 MMP1，且都是被調控基因，因此可推論其表現易於正常細胞是癌化後的結果，而非致癌的原因。至於肺腺癌特徵蛋白質交互作用網路，因為沒有上下游關係，因此只能推論，在癌細胞中，這些交互作用較正常細胞來得強或弱。我們亦計算了這些網路的基本拓樸性質，下表是肺腺癌特徵生物網路的基本拓樸特性：

	<Degree>	<Betweenness>	<Closeness>
肺腺癌特徵基因調控網路	1.25	1.25	0.15
肺腺癌特徵蛋白質交互作用網路	1.56	1.51	0.05

除了基本的拓樸特性外，針對這些特徵網路，我們利用 MLC 演算法(Enright *et al.*, 2002)來計算其可能存在的次網路。MLC 演算法是模擬流量在網路中的隨機流動過程，並產生出可能存在的次網路結構，這些次網路可能是癌症基因調控模組或者是蛋白質交互作用模組。以肺腺癌特徵蛋白質交互作用網路為例，其次網路結構如下：



三、癌症特徵次網路與癌症的關連性

若以 S 表示某個次網路的基因群，其中包含 k 個基因，以 G 表示以次序結構變異所篩檢出來的癌症相關基因群，假設此微陣列總共包含 y 個基因，而細胞內全部基因數為 N，令 I=S∩G 的基因數，則對 S 與 G 這兩群基因的分佈做關連分析(Haverty *et al.*, 2004)，可計算出 p-value：

$$p_{val} = \sum_{i=I}^{k} \frac{C_i^y C_{k-i}^{N-y}}{C_k^N}$$

p_{value} 越小，表示此癌症 G 與次網路 S 有較高的關連性，我們定義其關連強度為 $1-p_{value}$

我們計算了癌症相關的次網路以及其關連強度，結果發現其關連度往往不具統計上的

顯著性。

四、癌症相關生化網路資料庫

我們總共分析了 18 大類型、涵蓋 143 種亞型的癌症，並建立其相關生化網路資料庫(http://210.70.82.119/Cancer_Network/)，包含各類型癌症的特徵基因與其蛋白質產物、基因與蛋白質產物的註解、癌症特徵生物網路、生物次網路、生物次網路與癌症的關連性以及次網路之間可能的交互作用，這些資料在未來仍有待進一步分析與探討。

計畫成果自評

本計畫為兩年之計畫，在計畫進行過程中，教育學生具備一定之能力參與計畫，是最困難的一環。本計畫的第一年主要在整理生物網路資料庫以及蒐集癌症相關的微陣列實驗數據，同時教育學生培養參與計畫的能力，第二年則開始進入計畫的主要部分--癌症相關生物網路探勘，由於人類相關的生物網路資料仍不豐富，因此，最後的結果仍有很大的發展空間，相信隨著公用資料庫的資料量增加，未來應能有很大的改善，另一個問題是，癌症特徵基因的選取，以次序結構變異取代統計檢定，雖然可以得到較多的基因數，這在生物網路建立上是很有幫助的，但是，為了避免可能的偽陽性，我們提高了篩選的標準(p-value < 0.01)。最後；我們利用 MLC 演算法定義出基因調控次網路以及蛋白質交互作用次網路，代謝次網路則依據 KEGG 的分類法，並篩選出癌症相關的次網路，關於次網路的交互作用，我們往往得到無交互作用的結果，可能原因是建立的網路並不夠完整以及人類生化網路（除了代謝網路外）基因數太少，未來應持續探討。

本計畫已初步完成癌症相關生物網物建立與分析，有待未來進一步探討，個人相信以綜觀所有類型癌症的角度來看待癌症，而非僅僅探討個別癌症，在未來應該能有一些新的發現。

參考文獻

Aaboe, M., Birkenkamp-Demtroder, K., Wiuf, C., Sørensen, F.B., Thykjaer, T., Sauter, G., Jensen, K.M., Dyrskjøt, L., and Ørntoft, T. (2006) SOX4 expression in bladder carcinoma: clinical aspects and in vitro functional characterization. *Cancer Res.*, **66**, 3434-3442.

Albert, R. and Barabasi, A.L. (2005) Statistical mechanics of complex networks. *Reviews of modern physics*, **74**, 47-97.

Andersson, A., Ritz, C., Lindgren, D., Edén, P., Lassen, C., Heldrup, J., Olofsson, T., Råde, J., and Fontes, M., Porwit-Macdonald, A., Behrendtz, M., Höglund, M., Johansson, B., and Fioretos, T. (2007) Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia*, **21**, 1198-203.

Antonov, A., Tetko, I.V., and Mewes, H.W. (2006) A systematic approach to infer biological relevance and biases of gene network structures. *Nucleic Acids Research*, **34**, e6

Bader, G.D., Cary, M.P., and Sander, C. (2006) Pathguide: a Pathway Resource List. *Nucleic Acids Research*, **34**, D504-D506.

Ball, C.A., Demeter, J., Gollub, G., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Nitzberg, M., Wymore, F., Brown, P.O.1,, Sherlock, G. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Research*, **33**, D580-D582.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet.*, **37**, 382-90.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S., Du, J., Kau, T., Thomas, R.K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R.M., Demichelis, F., Hatton, C., Rubin, M.A., Garraway, L.A., Nelson, S.F., Liau, L., Mischel, P.S., Cloughesy, T.F., Meyerson, M., Golub, T.A., Lander, E.S., Mellinghoff, I.K., and Sellers, W.R. (2007) *Proc Natl Acad Sci U S A*, **104**, 20007-20012.

Beroukhim, R., Brunet, J.P., Di Napoli, A., Mertz, K.D., Seeley, A., Pires, M.M., Linhart, D., Worrell, R.A., Moch, H., Rubin, M.A., Sellers, W.R., Meyerson, M., Linehan, W.M., Kaelin, W.G. Jr., and Signoretti, S. (2009) Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res.*, **69**, 4674-4681.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., and Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A*, **98**, 13790-13795.

Bommert, K., Bargou, R.C., and Stuhmer, T. (2006) Signalling and survival pathways in multiple myeloma. *Eur. J. Cancer,* **42**, 1574-80

Buchholz, M., Braun, M., Heidenblut, A., Kestler. H.A., Klöppel, G., Schmiegel, W., Hahn, S.A., Lüttges, J., and Gress, T.M. (2005) Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions. *Oncogene*, **24**, 6626-6636.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167-171.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Garcia Lara, G., Oezcimen, A., Sansone, S., Rocca-Serra, P. (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **33**, D553-D555.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167-171.

Chambers, A., Angulo, A., Amaratunga, D., Guo, H., Jiang, Y., Wan, J.S., Bittner, A., Frueh, K., Jackson, M.R., Peterson, P.A., Erlander, M.G., and Ghazal, P. (1999) DNA microarrays of the complex human cytomegalovirus genome: Profiling kinetic class with drug sensitive viral gene expression. *J. Virol.*, **73**, 5757-5766.

Chen, H.C., Lee, H.C., Lin, T.Y., Li, W.H., and Chen, B.S. (2004) Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics*, **20,** 1914-1927.

Chen, J.J., Lin, Y.C., Yao, P.L., Yuan, A., Chen, H.Y., Shun, C.T., Tsai, M.F., Chen, C.H. and Yang, P.C. (2005) Tumor-associated macrophages: the double-edged sword in cancer progression. *J. Clin. Oncol.* **23**, 953-964.

Chen, J.J.W., Peck, K., Hong, T.M., Yang, S.C., Sher, Y.P., Shih, J.Y., Wu, R., Wu, C.W. and Yang, P.C. (2001) Global analysis of gene expression in invasion by a lung cancer model. *Cancer Res.* **61**, 5223-5230.

Chen, J., Zhao, P., Massaro, D., Clerch, L.B., Almon, R.R., DuBois, D.C., Jusko, W.J., and Hoffman, E.P. (2004) The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface. *Nucleic Acids Research*, **32**, D578-D581.

Chen, X., Leung, S.Y., Yuen, S.T., Chu, K.M., Ji, J., Li, R., Chan, A.S., Law, S., Troyanskaya, O.G., Wong, J., So, S., Botstein, D., and Brown, P.O. (2003) Variation in gene expression patterns in human gastric cancers. *Mol Biol Cell*, **14**, 3208-3215.

Chen, X., Cheung, S.T., So, S., Fan, S.T., Barry, C., Higgins, J., Lai, K.M., Ji, J., Dudoit, S., Ng, I.O., Van De Rijn, M., Botstein, D., and Brown, P.O. (2002) Gene expression patterns in human liver cancers. *Mol Biol Cell*, **13**, 1929-1939.

Chen, Y., Dougherty, E.D., and Bittner, M.L. (1997) Ratio-based decision and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.*, **2**, 364-374

Choi, Y.L,, Tsukasaki, K., O'Neill, M.C., Yamada, Y., Onimaru, Y., Matsumoto, K., Ohashi, J., Yamashita, Y., Tsutsumi, S., Kaneda, R., Takada, S., Aburatani, H., Kamihira, S., Nakamura, T., Tomonaga, M., and Mano, H. (2007) A genomic analysis of adult T-cell leukemia. *Oncogene*, **26**, 1245-1255.

Cutcliffe, C., Kersey, D., Huang, C.C., Zeng, Y., Walterhouse, D., and Perlman, E.J. (2005) Clear cell sarcoma of the kidney: up-regulation of neural markers with activation of the sonic hedgehog and Akt pathways. *Clin Cancer Res.*, **11**, 7986-7994.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Hernandez-Boussard, T., Cherry, J.M., Brown, P.O., Alizadeh, A.A. (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research*, **31**, 219-223.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.

Detwiller, K.Y., Fernando, N.T., Segal, N.H., Ryeom, S.W., D'Amore, P.A., and Yoon, S.S. (2005) Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A. *Cancer Res.*, **65**, 5881-5889.

Dorogovtsev, S. N. and Mendes, J. F. F. (2002) Evolution of networks. *Adv. Phys.*, **51**, 1079–1187.

Dürig, J., Bug, S., Klein-Hitpass, L., Boes, T., Jöns, T., Martin-Subero, J.I., Harder, L., Baudis, M., Dührsen, U., and Siebert, R. (2007) Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances,

mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia*, **21**, 2153-2163.

Dyrskjøt, L., Kruhøffer, M., Thykjaer, T., Marcussen, N., Jensen, J.L., Møller, K., and Ørntoft, T.F. (2004) Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res.*, **64**, 4040-4048.

Edgar, R., Domrachev, M., and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207-210.

Elfilali, A., Lair, S., Verbeke, C., La Rosa, P., Radvanyi, F. and Barillot, E. (2006) ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. *Nucleic Acids Research*, **34**, D613-D616.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, **30**, 1575-1584..

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., and Park, M. (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat Med.*, **14**, 518-527.

French, P.J., Swagemakers, S.M., Nagel, J.H., Kouwenhoven, M.C., Brouwer, E., van der Spek, P., Luider, T,M,, Kros, J.M., van den Bent, M.J., and Sillevis Smitt, P.A. (2005) Gene expression profiles associated with treatment response in oligodendrogliomas. *Cancer Res.*, **65**, 11335-11344.

Frierson, H.F., Jr, El-Naggar, A.K., Welsh, J.B., Sapinoso, L.M., Su, A.I., Cheng, J., Saku, T., Moskaluk, C.A., Hampton, G.M. (2002) Large scale molecular analysis identifies genes with altered expression in salivary adenoid cystic carcinoma. *Am J Pathol.*, **161**, 1315-1323.

Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D., and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*, **98**, 13784-13789.

Gaspar, C., Cardoso, J., Franken, P., Molenaar, L., Morreau, H., Möslein, G., Sampson, J., Boer, J.M., de Menezes, R.X., and Fodde, R. (2008) Cross-species comparison of human and mouse intestinal polyps reveals conserved mechanisms in adenomatous polyposis coli (APC)-driven tumorigenesis. *Am J Pathol.*, **172**, 1363-80.

Gisiger, T. (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism?

Biol. Rev., **76**, 161-209.
Ginos, M.A., Page, G.P., Michalowicz, B.S., Patel, K.J., Volker, S.E., Pambuccian, S.E., Ondrey, F.G., Adams, G.L., and Gaffney, P.M. (2004) Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer Res.*, **64**, 55-63.

Giordano, T.J., Thomas, D.G., Kuick, R., Lizyness, M., Misek, D.E., Smith, A.L., Sanders, D., Aljundi, R.T., Gauger, P.G., Thompson, N.W., Taylor, J.M., and Hanash, S.M. (2003) Distinct transcriptional profiles of adrenocortical tumors uncovered by DNA microarray analysis. *Am J Pathol.*, **162**, 521-531.

Gordon, G.J., *et al*. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, **62**, 4963-4967.

Gordon, G.J., Rockwell, G.N., Jensen, R.V., Rheinwald, J.G., Glickman, J.N., Aronson, J.P., Pottorf, B.J., Nitz, M.D., Richards, W.G., Sugarbaker, D.J., and Bueno, R. (2005) Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *Am J Pathol.*, **166**, 1827-1840.

Graudens, E., Boulanger, V., Mollard, C., Mariage-Samson, R., Barlet, X., Grémy, G., Couillault, C., Lajémi, M., Piatier-Tonneau, D., Zaborski, P., Eveno, E., Auffray, C., and Imbeaud, S. (2006) Deciphering cellular states of innate tumor drug responses. *Genome Biol.*, **7**, R19.

Gutmann, D.H., Hedrick, N.M., Li, J., Nagarajan, R., Perry, A., and Watson, M.A. (2002) Comparative gene expression profile analysis of neurofibromatosis 1-associated and sporadic pilocytic astrocytomas. *Cancer Res.*, **62**, 2085-2091.

Hao, Y., Triadafilopoulos, G., Sahbaie, P., Young, H.S., Omary, M.B., and Lowe, A.W. (2006) Gene expression profiling reveals stromal genes expressed in common between Barrett's esophagus and adenocarcinoma. *Gastroenterology*, **131**, 925-933.

Haqq, C., Nosrati, M., Sudilovsky, D., Crothers, J., Khodabakhsh, D., Pulliam, B.L., Federman, S., Miller, J.R., Allen, R.E., Singer, M.I., Leong, S.P,, Ljung, B.M., Sagebiel, R.W., and Kashani-Sabet, M. (2005) The gene expression signatures of melanoma progression. *Proc Natl Acad Sci U S A*, **102**, 6092-6097.

Haslinger, C., Schweifer, N., Stilgenbauer, S., Döhner, H., Lichter, P., Kraut, N., Stratowa, C., and Abseher, R. (2004) Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol.*, **22**, 3937-3949.

Haverty, P.M., Hansen, U., and Weng, Z. (2004) Computational inference of transcriptional regulatory networks from expression proÆling and transcription factor binding site identification. *Nucleic Acids Research,* **32**, 179-188.

Haverty, P., Weng, Z., Best, N., Auerback, K., Hsiao, L., Jensen, R., Gullans, S. (2002) HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Research,* **30**, 214-217

Huang, Y., Prasad, M., Lemon, W.J., Hampel, H., Wright, F.A., Kornacker, K., LiVolsi, V., Frankel, W., Kloos, R.T., Eng, C., Pellegata, N.S., and de la Chapelle, A. (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc Natl Acad Sci U S A*, **98**, 15044-15049.

Iacobuzio-Donahue, C.A., Maitra, A., Olsen, M., Lowe, A.W., van Heek, N.T., Rosty, C., Walter, K., Sato, N., Parker, A., Ashfaq, R., Jaffee, E., Ryu, B., Jones, J., Eshleman, J.R., Yeo, C.J., Cameron, J.L., Kern, S.E., Hruban, R.H., Brown, P.O., and Goggins, M. (2003) Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am J Pathol.*, **162**, 1151-1162.

Ishikawa, M., Yoshida, K., Yamashita, Y., Ota, J., Takada, S., Kisanuki, H., Koinuma, K., Choi, Y.L., Kaneda, R., Iwao, T., Tamada, K., Sugano, K., and Mano, H. (2005) Experimental trial for diagnosis of pancreatic ductal carcinoma based on gene expression profiles of pancreatic ductal cells. *Cancer Sci.*, **96**, 387-393.

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature,* **411**, 41–41.

Kaiser, S., Park, Y.K., Franklin, J.L., Halberg, R.B., Yu, M., Jessen, W.J., Freudenberg, J., Chen, X., Haigis, K., Jegga, A.G., Kong, S., Sakthivel, B., Xu, H., Reichling, T., Azhar, M., Boivin, G.P., Roberts, R.B., Bissahoyo, A.C., Gonzales, F., Bloom, G.C., Eschrich, S., Carter, S.L., Aronow, J.E., Kleimeyer, J., Kleimeyer, M., Ramaswamy, V., Settle, S.H., Boone, B., Levy, S., Graff, J.M., Doetschman, T., Groden, J., Dove, W.F., Threadgill, D.W., Yeatman, T.J., Coffey, R.J. and Jr, Aronow, B.J. (2007) Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. *Genome Biol.*, **8**, R131.

Katoh, Y., and Katoh, M. (2005) Hedgehog signaling pathway and gastric cancer. *Cancer Biol. Ther.*, **4**, 1050-1054.

Keller, J., Gader, P., and Hocaoglu, A.K. (2000) Fuzzy Integrals in Image Processing and Recognition. *Fuzzy Measures and Integrals: Theory and Applications*, pp. 435-466.

Ki, D.H., Jeung, H.C., Park, C.H., Kang, S.H., Lee, G.Y., Lee, W.S., Kim, N.K., Chung, H.C., and Rha, S.Y. (2007) Whole genome analysis for liver metastasis gene signatures in colorectal cancer. *Int J Cancer*, **121**, 2005-2012.

Korkola, J.E., Houldsworth, J., Chadalavada, R.S., Olshen, A.B., Dobrzynski, D., Reuter, V.E., Bosl, G.J., and Chaganti, R.S. (2006) Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13.31, is associated with in vivo differentiation of human male germ cell tumors. *Cancer Res.*, **66**, 820-827.

Lee, J., Kotliarova, S., Kotliarov, Y., Li, A., Su, Q., Donin, N.M., Pastorino, S., Purow, B.W., Christopher, N., Zhang, W., Park, J.K., and Fine, H.A. (2006) Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell*, **9**, 391-403.

Liang, Y., Diehn, M., Watson, N., Bollen, A.W., Aldape, K.D., Nicholas, M.K., Lamborn, K.R., Berger, M.S., Botstein, D., Brown, P.O., and Israel, M.A. (2005) Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A*, **102**, 5814-5819.

Liao, C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., and Roychowdhury, V.P. (2005) Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Nat. Acad. Sci.*, **100**, 15522–15527.

Liu, C.C., Chen, W.S, Lin, C.C., Liu, H.S., Yang, P.C., Chang, P.C., and Chen, J.W. (2006) Topology-based cancer classification and related pathway mining using microarray data. *Nucleic Acids Research*, **34**, 4069-4080.

Logsdon, C.D., Simeone, D.M., Binkley, C., Arumugam, T., Greenson, J.K., Giordano, T.J., Misek, D.E., Kuick, R., and Hanash, S. (2003) Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer. *Cancer Res.*, **63**, 2649-2657.

Lu, K.H., Patterson, A.P., Wang, L., Marquez, R.T., Atkinson, E.N., Baggerly, K.A., Ramoth, L.R., Rosen, D.G., Liu, J., Hellstrom, I., Smith, D., Hartmann, L., Fishman, D., Berchuck, A., Schmandt, R., Whitaker, R., Gershenson, D.M., Mills, G.B., Bast. R.C. (2004) Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis. *Clin Cancer Res.*, **10**, 3291-3300.

Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F., and Trajanoski, Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological

pathways. *Nucleic Acids Research,* **33**, W633-W637.

Morrison, C., Farrar, W., Kneile, J., Williams, N., Liu-Stratton, Y., Bakaletz, A., Aldred, M.A., and Eng, C. (2004) Molecular classification of parathyroid neoplasia by gene expression profiling. *Am J Pathol.*, **165**, 565-576.

Mutter, G.L., Baak, J.P., Fitzgerald, J.T., Gray, R., Neuberg, D., Kust, G.A., Gentleman, R., Gullans, S.R., Wei, L.J., and Wilcox, M. (2001) Global expression changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation. *Gynecol Oncol.*, **83**, 177-185.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., and Golub, T.R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature,* **415**, 436-442.

Popescu, M., Keller, J.M., and Mitchell, J.A. (2006) Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS*, **3**, 263-274.

Pyeon, D., Newton, M.A., Lambert, P.F., den Boon, J.A., Sengupta, S., Marsit, C.J., Woodworth, C.D., Connor, J.P., Haugen, T.H., Smith, E.M., Kelsey, K.T., Turek, L.P., and Ahlquist, P. (2007) Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.*, **67**, 4605-4619.

Quade, B.J., Wang, T.Y., Sornberger, K., Dal Cin, P., Mutter, G.L., and Morton, C.C. (2004) Molecular pathogenesis of uterine smooth muscle tumors from transcriptional profiling. *Genes Chromosomes Cancer*, **40**, 97-108.

Radvanyi, L., Singh-Sandhu, D., Gallichan, S., Lovitt, C., Pedyczak, A., Mallo, G., Gish, K., Kwok, K., Hanna, W., Zubovits, J., Armes, J., Venter, D., Hakimi, J., Shortreed, J., Donovan, M., Parrington, M., Dunn, P., Oomen, R., Tartaglia, J., and Berinstein, N.L. (2005) The gene associated with trichorhinophalangeal syndrome in humans is overexpressed in breast cancer. *Proc Natl Acad Sci U S A*, **102**, 11005-11010.

Rhodes, D.R. et al. (2004) OCOMINE: a cancer microarray database and integrated data mining platform. *Neoplasia,* **6**, 1-6.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T.R., Ghosh, D., and Chinnaiyan, A.M. (2005) Mining for regulatory programs in the cancer transcriptome. *Nature Genet.*, **37**,

579-583.

Richardson, A.L., Wang, Z.C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J.D., Livingston, D.M., and Ganesan, S. (2006) X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**, 121-132.

Riker, A.I., Enkemann, S.A., Fodstad, O., Liu, S., Ren, S., Morris, C., Xi, Y., Howell, P., Metge, B., Samant, R.S., Shevde, L.A., Li, W., Eschrich, S., Daud, A., Ju, J., and Matta, J. (2008) The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med Genomics*, **28**, 1-13.

Ryan, P.E., Davies, G.C., Nau, M.M., and Lipkowitz, S. (2006) Regulating the regulator: negative regulation of Cbl ubiquitin ligases. *Trends Biochem Sci.*, **31**, 79-88.

Sabates-Bellver, J., Van der Flier, L.G., de Palo, M., Cattaneo, E., Maake, C., Rehrauer, H., Laczko, E., Kurowski, M.A., Bujnicki, J.M., Menigatti, M., Luz, J., Ranalli, T.V., Gomes, V., Pastorelli, A., Faggiani, R., Anti, M., Jiricny, J., Clevers, H., and Marra, G. (2007) Transcriptome profile of human colorectal adenomas. *Mol Cancer Res.*, **5**, 1263-1275.

Santegoets, L.A., Seters, M., Helmerhorst, T.J., Heijmans-Antonissen, C., Hanifi-Moghaddam, P., Ewing, P.C., van Ijcken, W.F., van der Spek, P.J., van der Meijden, W.I., and Blok, L.J. (2007) HPV related VIN: highly proliferative and diminished responsiveness to extracellular signals. *Int J Cancer*, **121**, 759-766.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, **270**, 467-470.

Schlingemann, J., Habtemichael, N., Ittrich, C., Toedt, G., Kramer, H., Hambek, M., Knecht, R., Lichter, P., Stauber, R., and Hahn, M. (2005) Patient-based cross-platform comparison of oligonucleotide microarray expression profiles. *Lab Invest.*, **85**, 1024-1039.

Segara, D., Biankin, A.V., Kench, J.G., Langusch, C.C., Dawson, A.C., Skalicky, D.A., Gotley, D.C., Coleman, M.J., Sutherland, R.L., and Henshall, S.M. (2005) Expression of HOXB2, a retinoic acid signaling target in pancreatic cancer and pancreatic intraepithelial neoplasia. *Clin Cancer Res.*, **11**, 3587-3596.

Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liau, L.M., Cloughesy, T.F., Mischel, P.S., and Nelson, S.F. (2003) Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, **22**, 4918-4923.

Shmueli, O., Shmoish, M., Rosen, N., Benjamin-Rodrig, H., Yanai, I., Ophir, R., Shklar, M.,

Almashanu, L., Madi, A., Sirota, A., Kats, P., Chalifa-Caspi, V., Safran, M., Lancet, D. (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Research*, **31**, 142-146.

Sjöblom et al., (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268-274.

Skotheim, R.I., Lind, G.E., Monni, O., Nesland, J.M., Abeler, V.M., Fosså, S.D., Duale, N., Brunborg, G., Kallioniemi, O., Andrews, P.W., and Lothe, R.A. (2005) Differentiation of human embryonal carcinomas in vitro and in vivo reveals expression profiles relevant to normal development. *Cancer Res.*, **65**, 5588-5598.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Eystein Lønning, P., and Børresen-Dale, A.L. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, **98**, 10869-10874.

Storey,J.D., and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Strogatz, S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Storz, M.N., van de Rijn, M., Kim, Y.H., Mraz-Gernhard, S., Hoppe, R.T., and Kohler, S. (2003) Gene expression profiles of cutaneous B cell lymphoma. *J Invest Dermatol.*, **120**, 865-870.

Su, L.J., Chang, C.W., Wu, Y.C., Chen, K.C., Lin, C.J., Liang, S.C., Lin, C.H., Whang-Peng, J., Hsu, S.L., Chen, C.H., and Huang, C.Y. (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*, **1**, 140.

Sun, L., Hui, A.M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., Rosenblum, M., Mikkelsen, T., and Fine, H.A. (2006) Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, **9**, 287-300.

Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., and Wang, Y. (2005) Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin Cancer Res.*, **11**, 7234-7242.

Tomlins, S.A., Mehra, R., Rhodes, D.R., Cao, X., Wang, L., Dhanasekaran, S.M., Kalyana-Sundaram, S., Wei, J.T., Rubin, M.A., Pienta, K.J., Shah, R.B., and Chinnaiyan, A.M.

(2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet.*, **39**, 41-51.

Toruner, G.A., Ulger, C., Alkan, M., Galante, A.T., Rinaggio, J., Wilk, R., Tian, B., Soteropoulos, P., Hameed, M.R., Schwalb, M.N., and Dermody, J.J. (2004) Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. *Cancer Genet Cytogenet.*, **154**, 27-35.

Turashvili, G., Bouchal, J., Baumforth, K., Wei, W., Dziechciarkova, M., Ehrmann, J., Klein, J., Fridman, E., Skarda, J., Srovnal, J., Hajduch, M., Murray, P., and Kolar, Z. (2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer*, **27**, 7-55.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to ionizing radiation respone. *Proc. Nat. Acad. Sci.*, **98**, 5116-5121.

Valk, P.J., Verhaak, R.G., Beijen, M.A., Erpelinck, C.A., Barjesteh van Waalwijk van Doorn-Khosrovani, S., Boer, J.M., Beverloo, H.B., Moorhouse, M.J., van der Spek, P.J., Löwenberg, B., and Delwel, R. (2004) Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med.*, **350**, 1617-1628.

Vanaja, D.K., Cheville, J.C., Iturria, S.J., and Young, C.Y. (2003) Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. *Cancer Res.*, **63**, 3877-3882.

Watson, M.A., Gutmann, D.H., Peterson, K., Chicoine, M.R., Kleinschmidt-DeMasters, B.K., Brown, H.G., and Perry, A. (2002) Molecular characterization of human meningiomas by gene expression profiling using high-density oligonucleotide microarrays. *Am J Pathol.*, **161**, 665-672.

Whitehead, A., and Whitehead, J. (1991) A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, **10**, 1665-1677.

Winzeler, E.A., *et al*. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6.

Wurmbach, E., Chen, Y.B., Khitrov, G., Zhang, W., Roayaie, S., Schwartz, M., Fiel, I., Thung, S., Mazzaferro, V., Bruix, J., Bottinger, E., Friedman, S., Waxman, S., and Llovet, J.M. (2007) Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. *Hepatology*, **45**, 938-947.

Ye, H., Yu, T., Temam, S., Ziober, B.L., Wang, J., Schwartz, J.L., Mao, L., Wong, D.T., and Zhou,

X. (2008) Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics*, **6**, 69.

Yeoh, E.-J., *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell,* **1**, 133–143.

Yugi, K., Nakayama, Y., Kojima, S., Kitayama, T., and Tomita, M. (2005) A microarray data-based semi-kinetic method for predicting quantitative dynamics of genetic networks. *BMC Bioinformatics*, **6**, 299-315

Yusenko, M.V., Kuiper, R.P., Boethe, T., Ljungberg, B., van Kessel, A.G., and Kovacs, G. (2009) High-resolution DNA copy number and gene expression analyses distinguish chromophobe renal cell carcinomas and renal oncocytomas. *BMC Cancer*, **9**, 152.

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., Anaissie, E., Morris, C., Muwalla, F., van Rhee, F., Fassas, A., Crowley, J., Tricot, G., Barlogie, B., and Shaughnessy, J. (2002) Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood*, **99**, 1745-1757.

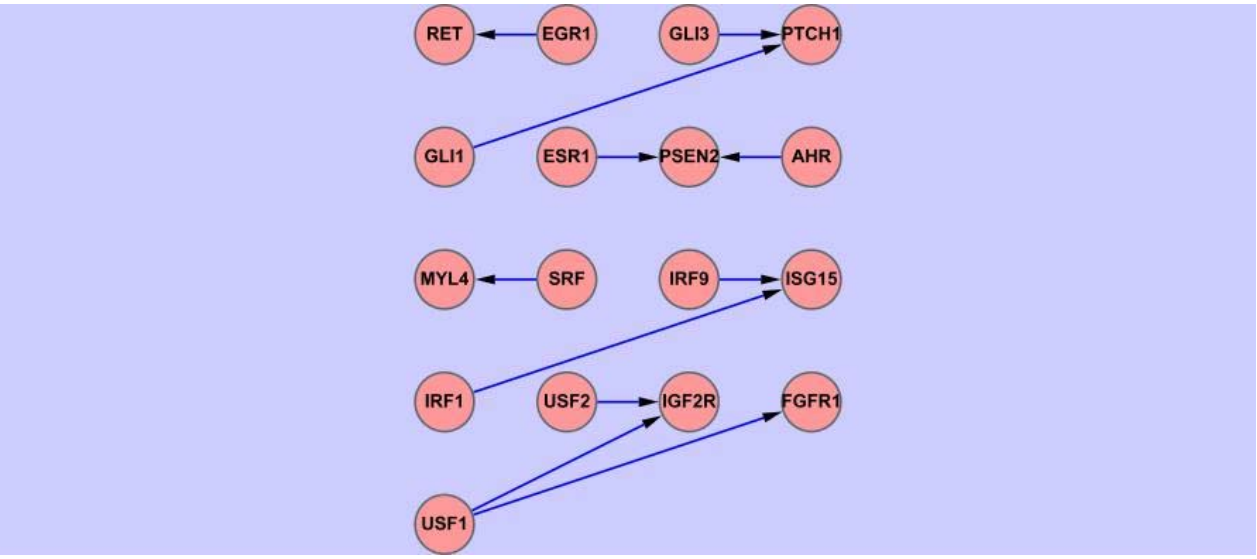
Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D.R., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Zangari, M., Dhodapkar, M., and Shaughnessy, J.D. (2007) Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis. *Blood*, **109**, 1692-1700.

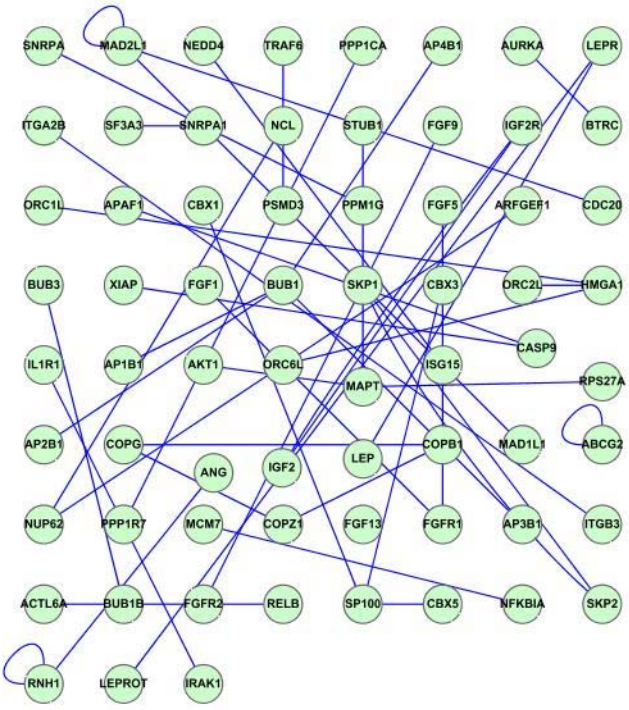
Zhao, H., Langerød, A., Ji, Y., Nowels, K.W., Nesland, J.M., Tibshirani, R., Bukholm, I.K., Kåresen, R., Botstein, D., Børresen-Dale, A.L., and Jeffrey, S.S. (2004) Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol Biol Cell*, **15**, 2523-2536.

Zou, T.T., Selaru, F.M., Xu, Y., Shustova, V., Yin, J., Mori, Y., Shibata, D., Sato, F., Wang, S., Olaru, A., Deacu, E., Liu, T.C., Abraham, J.M., and Meltzer, S.J. (2002) Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. *Oncogene,* **21**, 4855-4862.

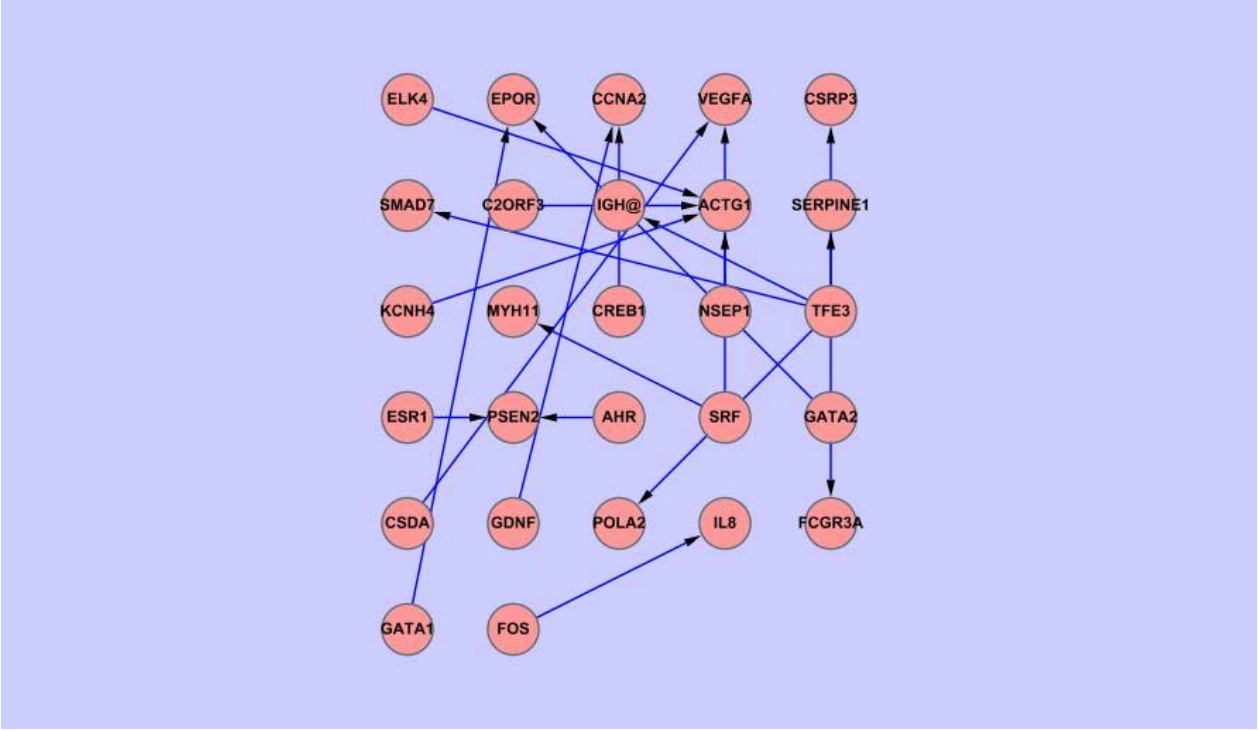
附錄一、常見癌症的特徵基因調控網路與特徵蛋白質交互作用網路

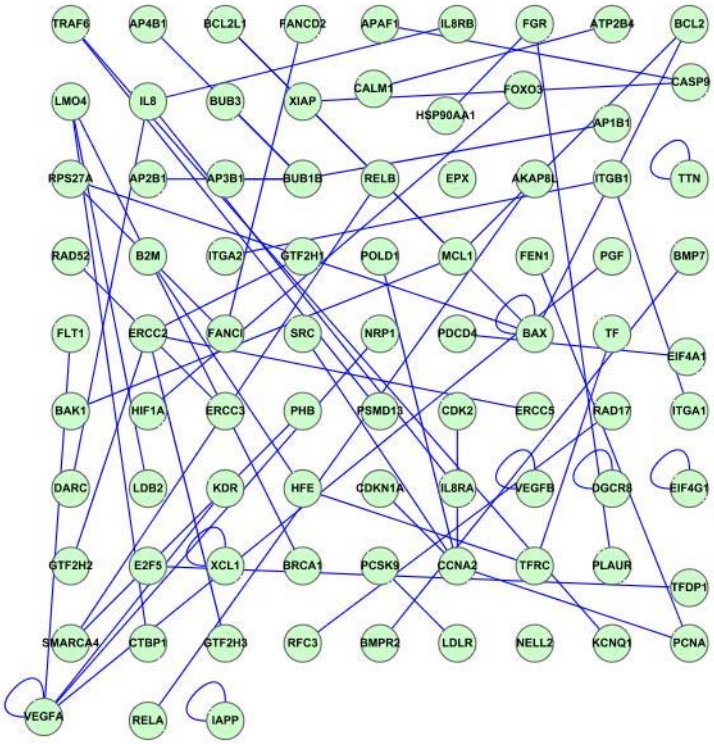
Type:Bladder Cancer　　Subtype:Infiltrating Bladder Urothelial Carcinoma



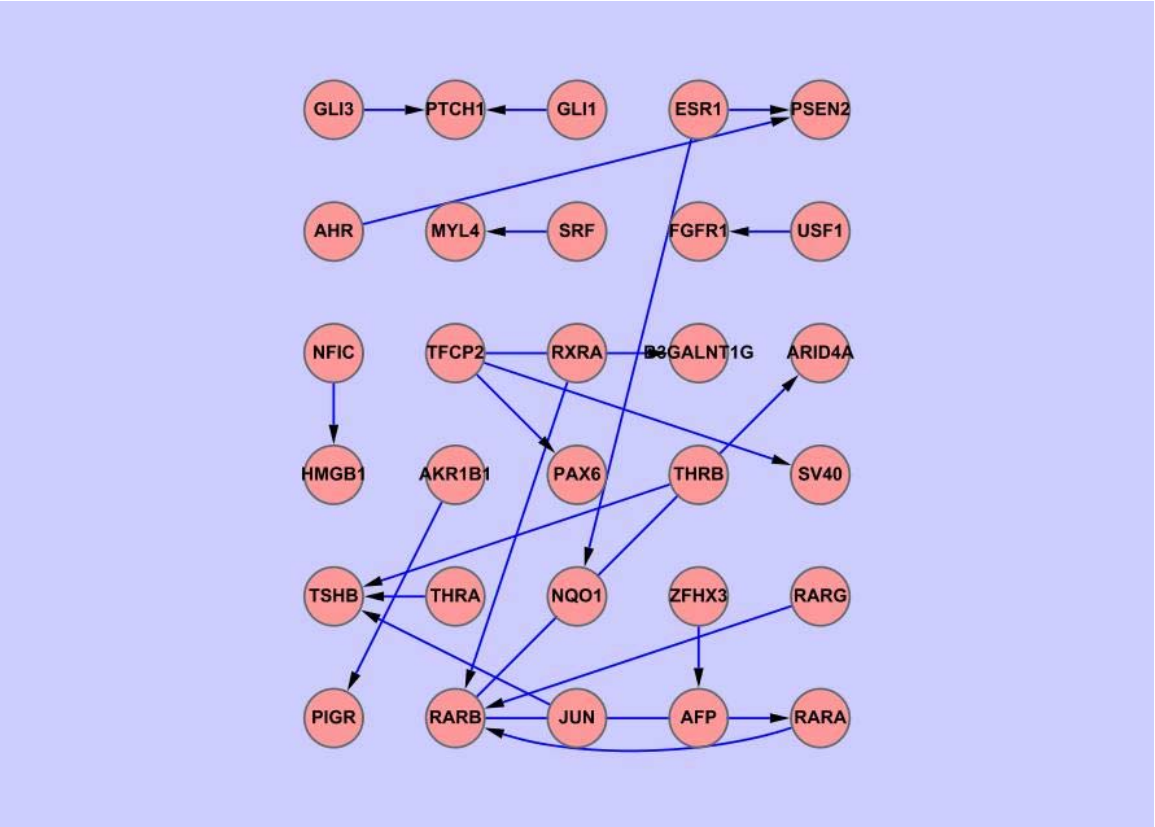


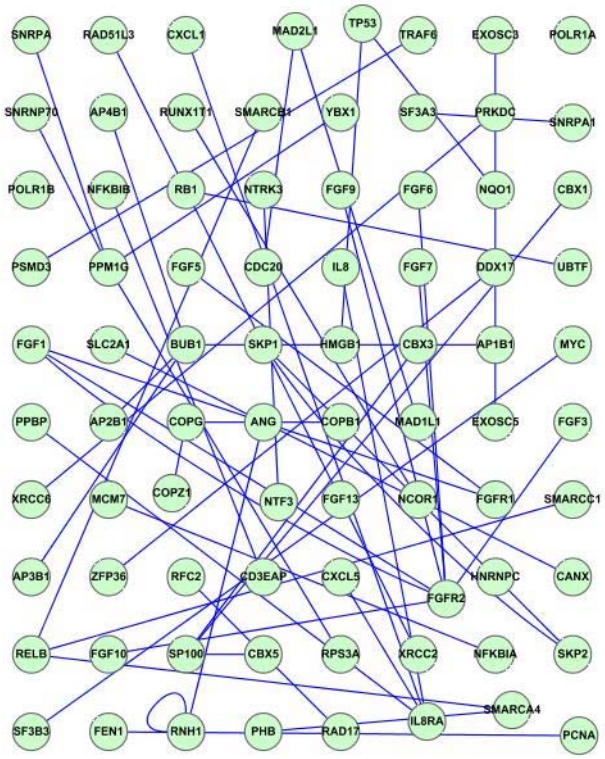
Type:Bladder Cancer Subtype:Stage 0is Bladder Urothelial Carcinoma



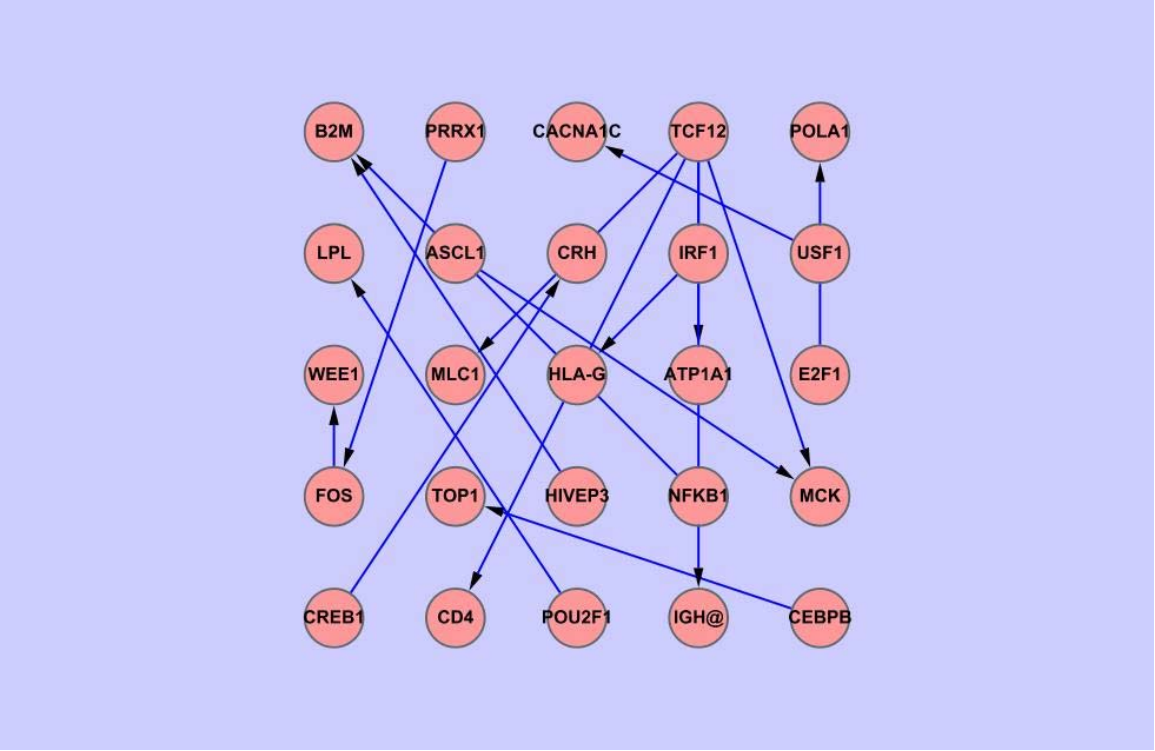


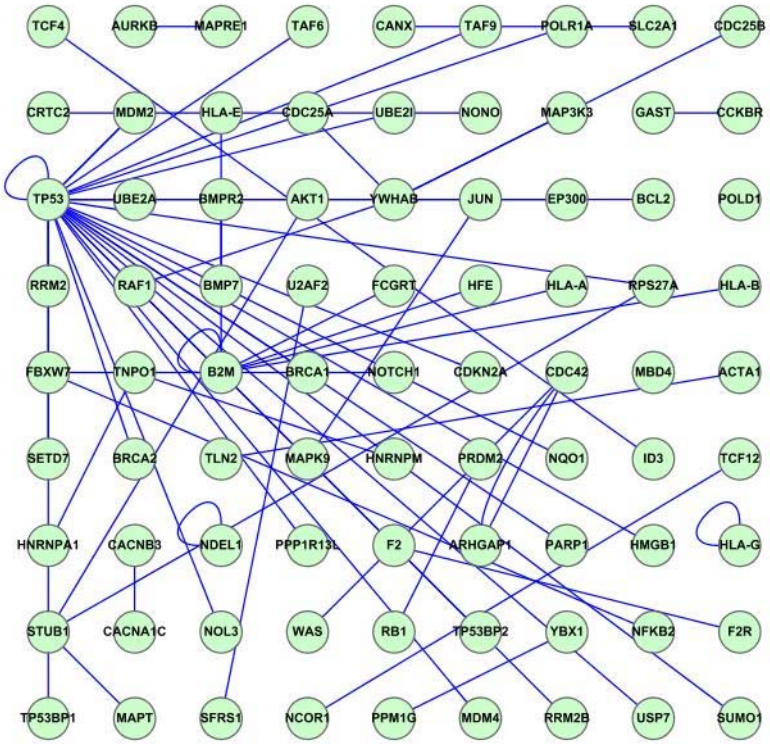
Type:Bladder Cancer Subtype:Superficial Bladder Cancer



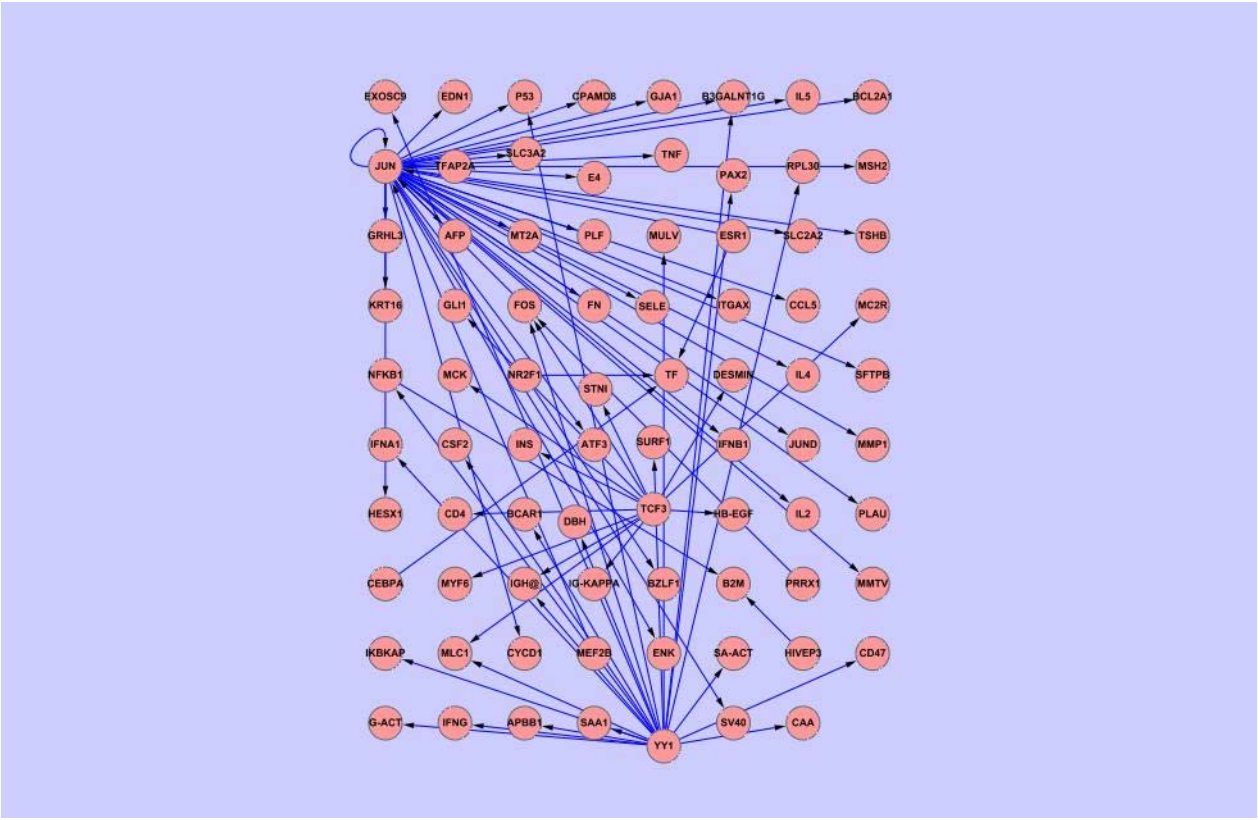


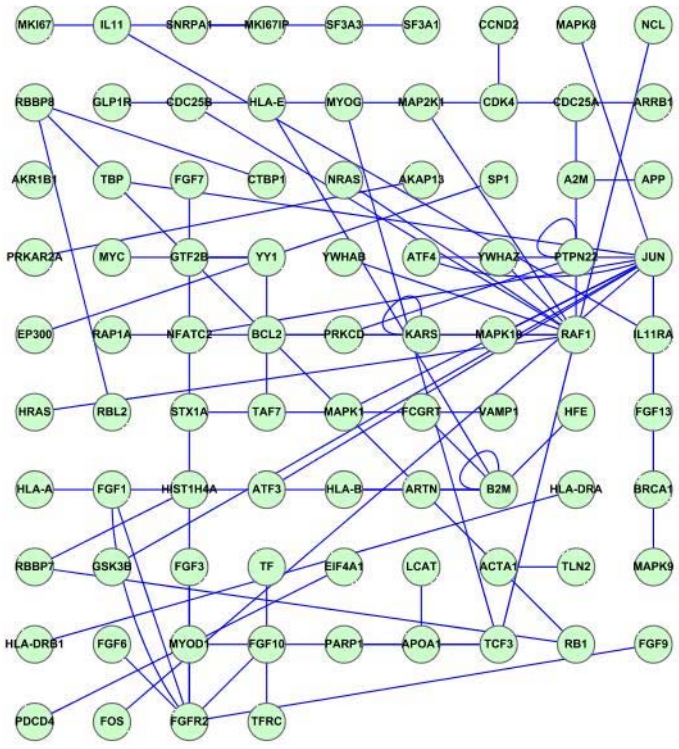
Type:Brain and CNS Cancer Subtype:Anaplastic Astrocytoma



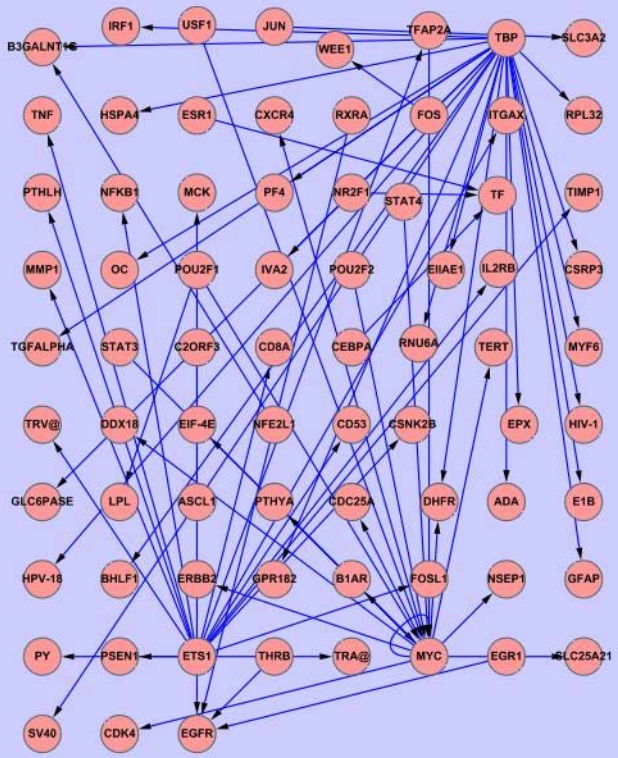


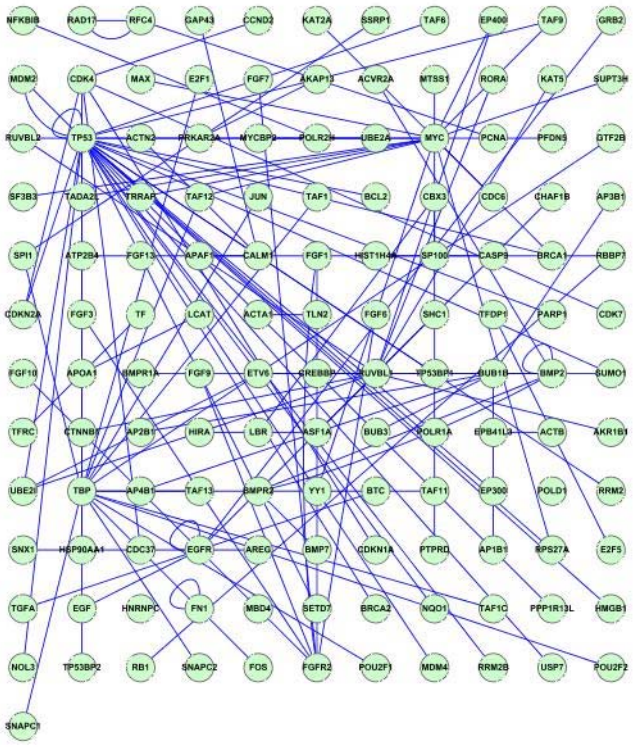
Type:Brain and CNS Cancer Subtype:Anaplastic Oligoastrocytoma



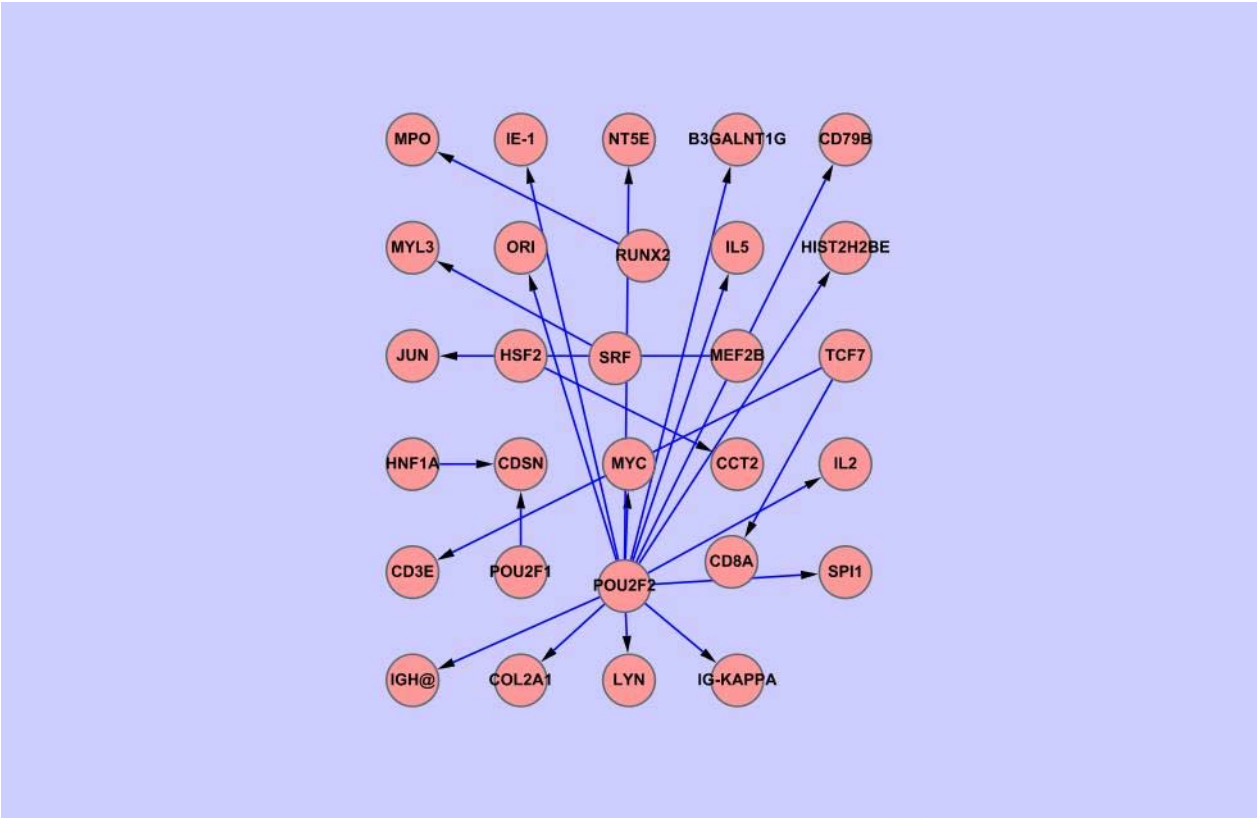


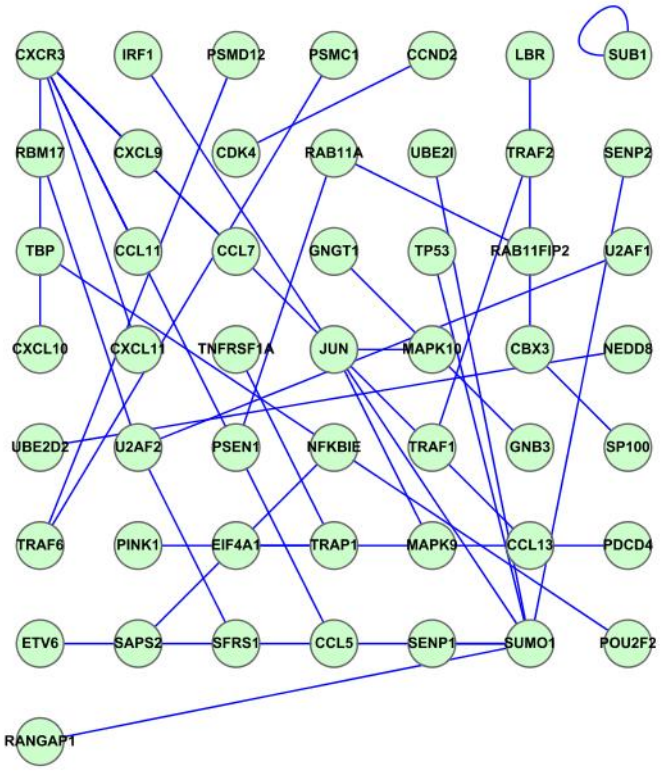
Type:Brain and CNS Cancer Subtype:Anaplastic Oligodendroglioma



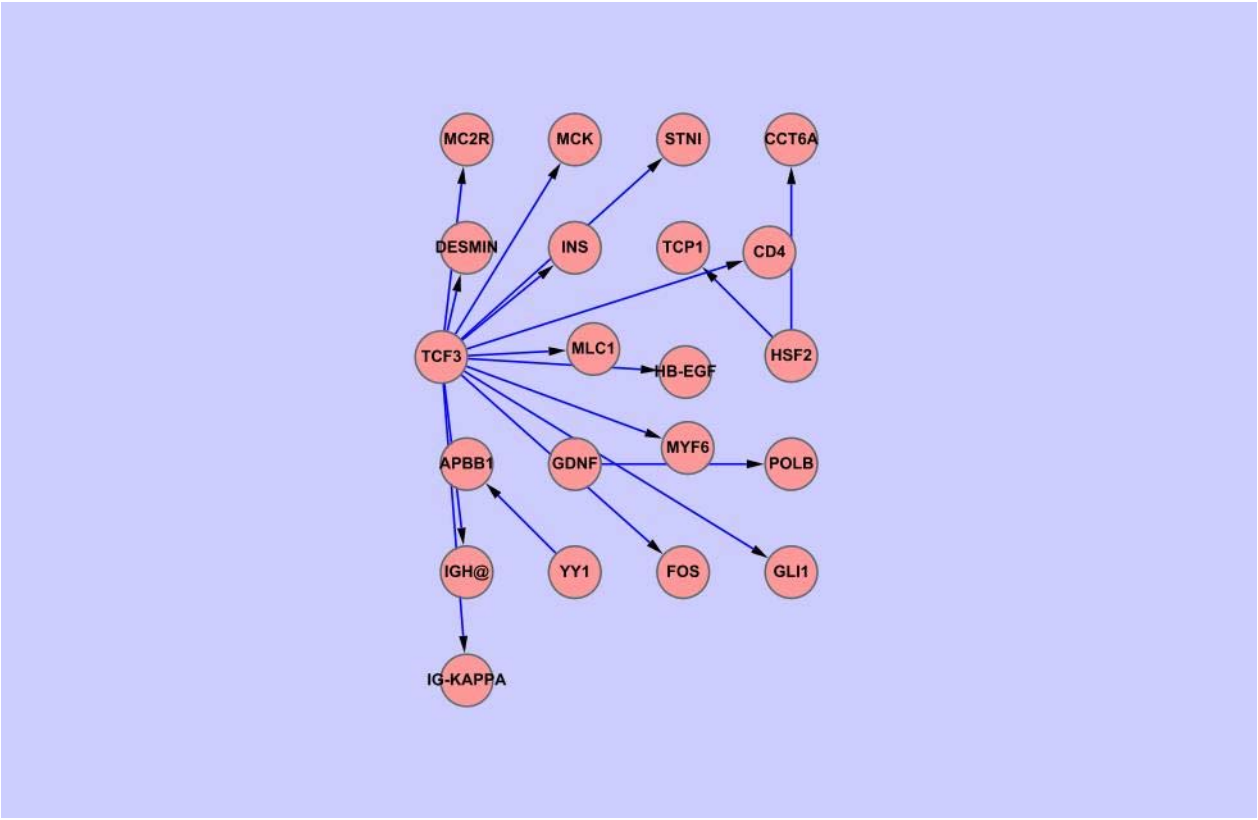


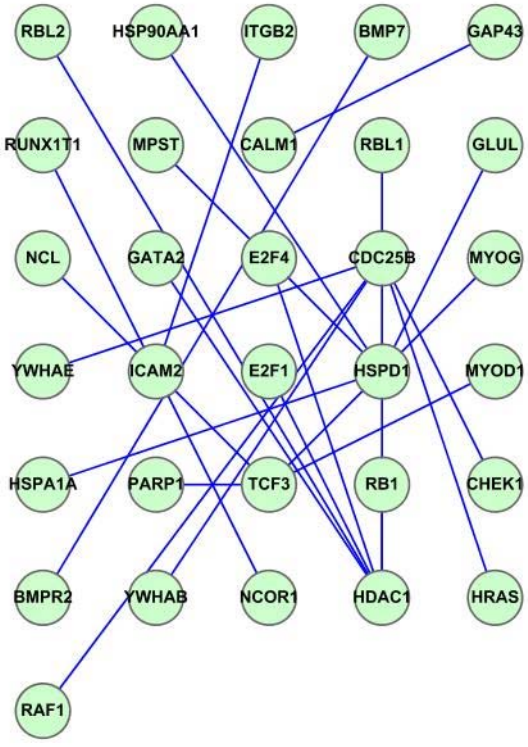
Type:Brain and CNS Cancer Subtype:Astrocytoma



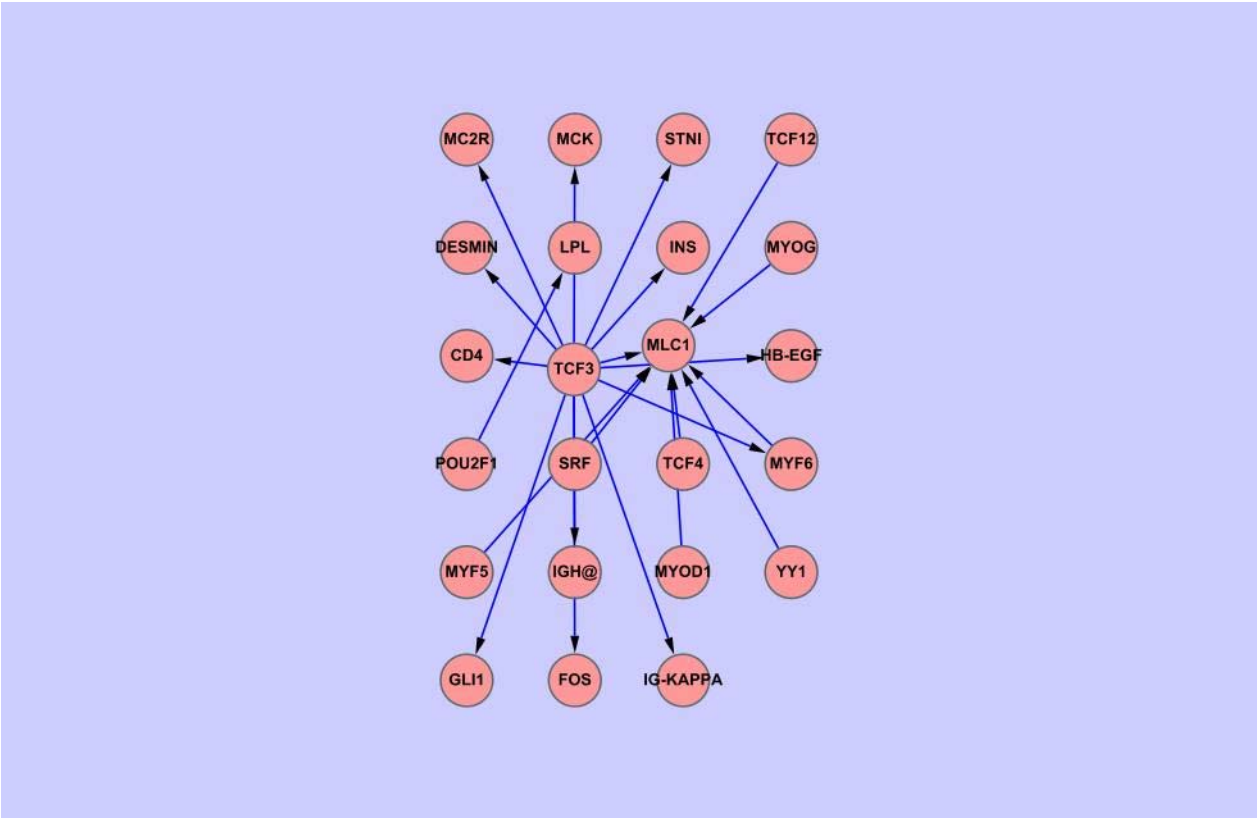


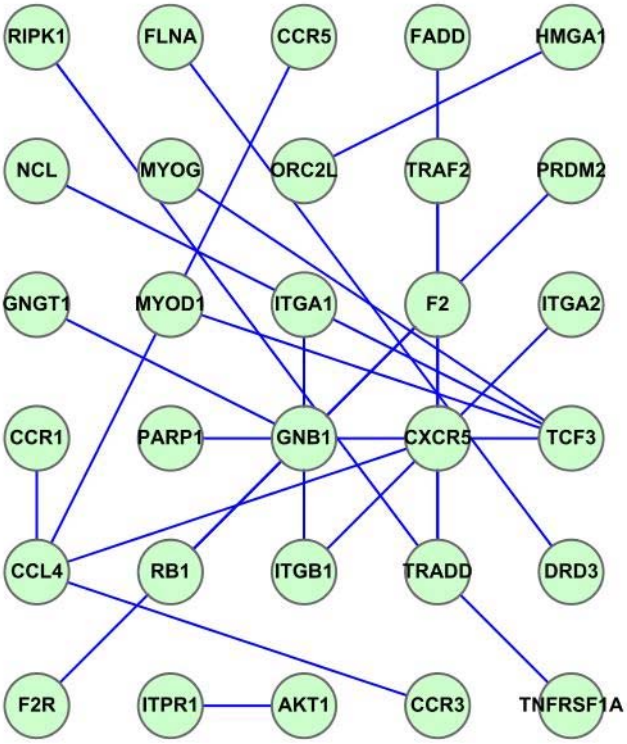
Type:Brain and CNS Cancer Subtype:Atypical Teratoid/Rhabdoid Tumor



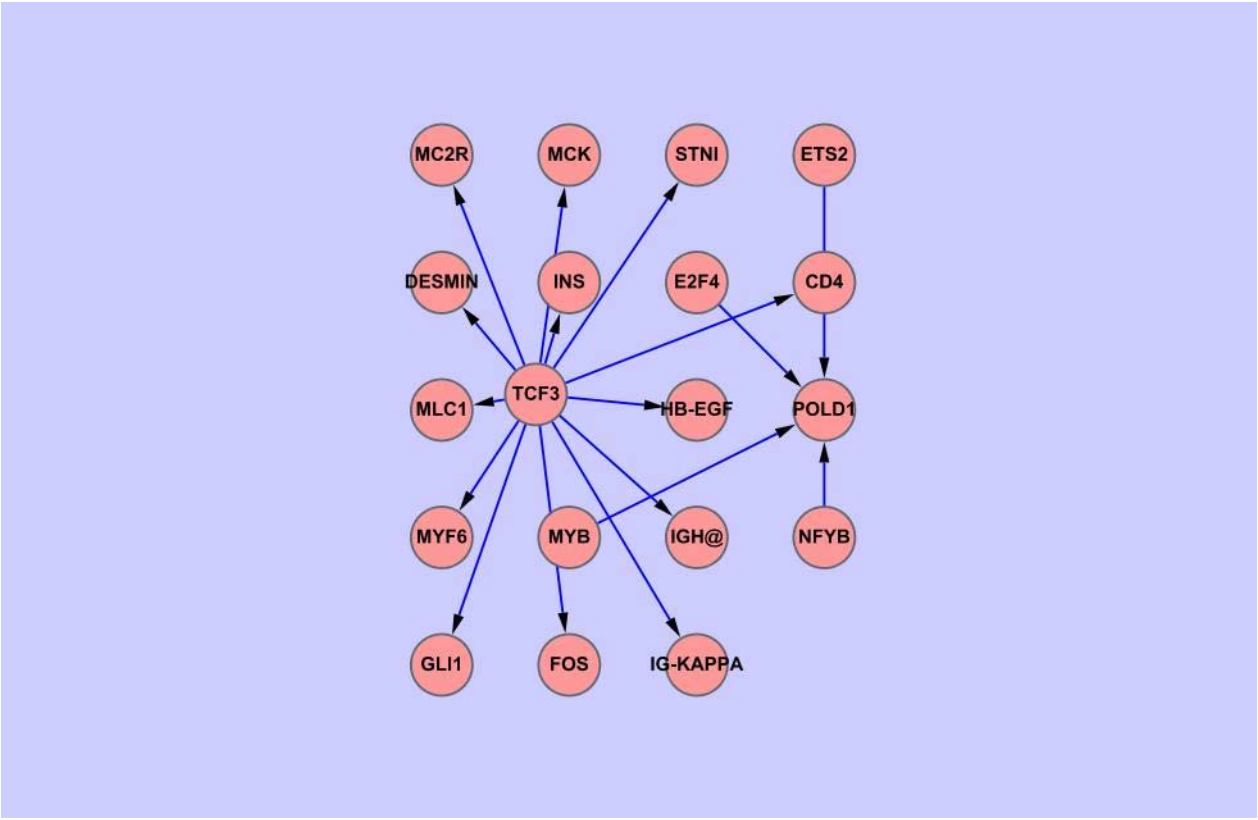


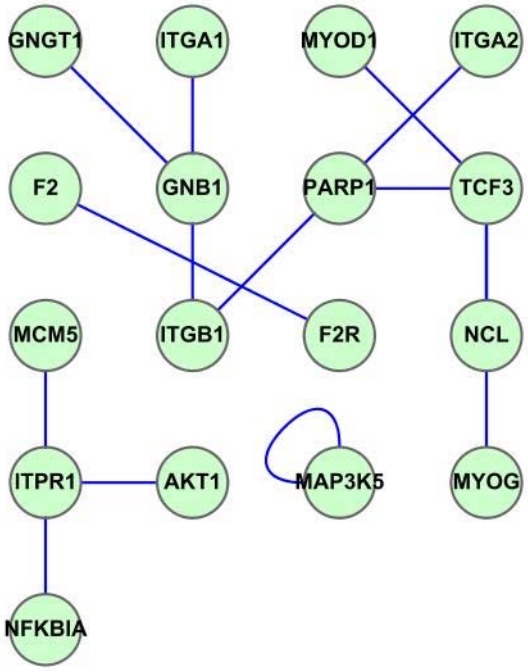
Type:Brain and CNS Cancer Subtype:Classic Medulloblastoma



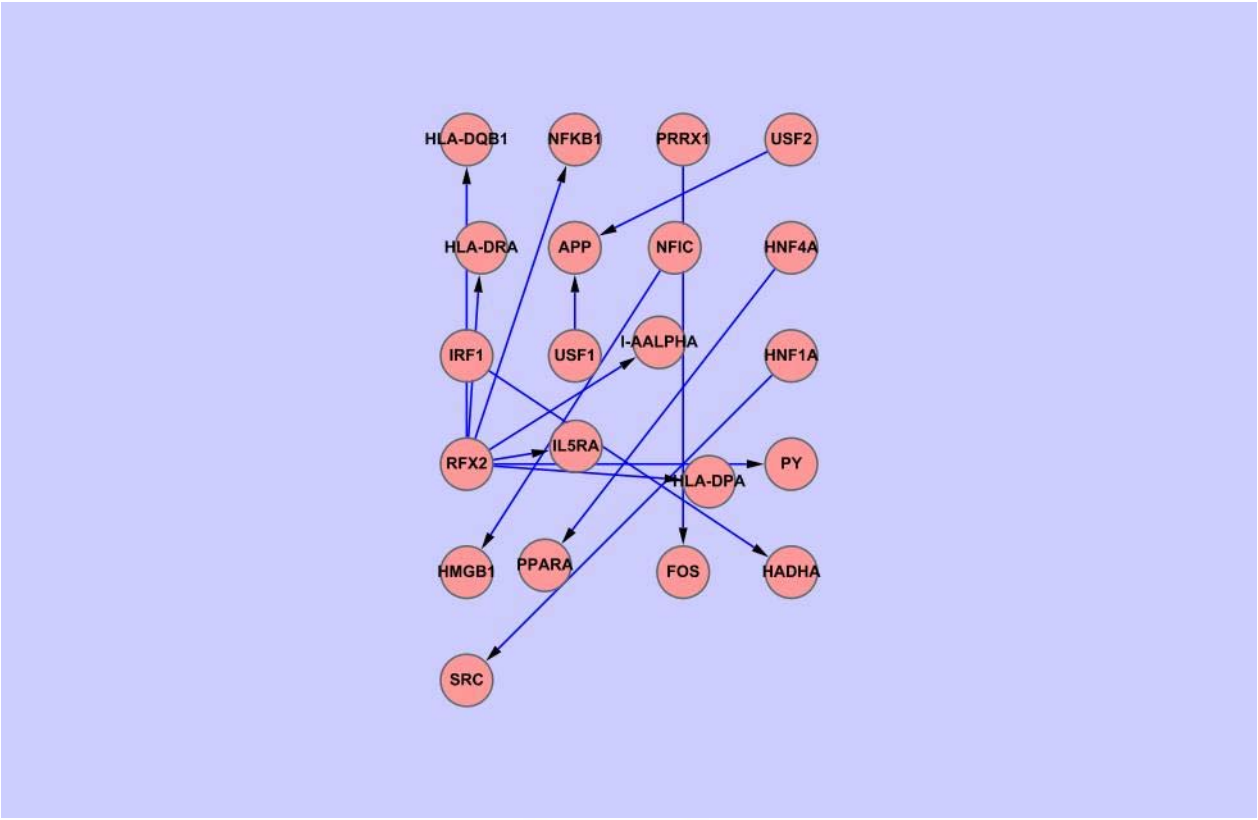


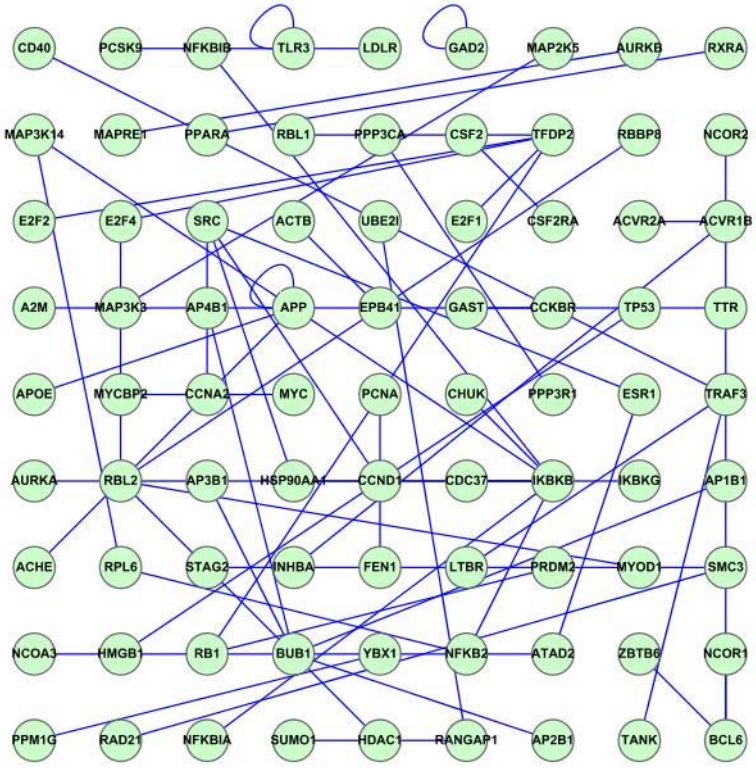
Type:Brain and CNS Cancer Subtype:Desmoplastic Medulloblastoma



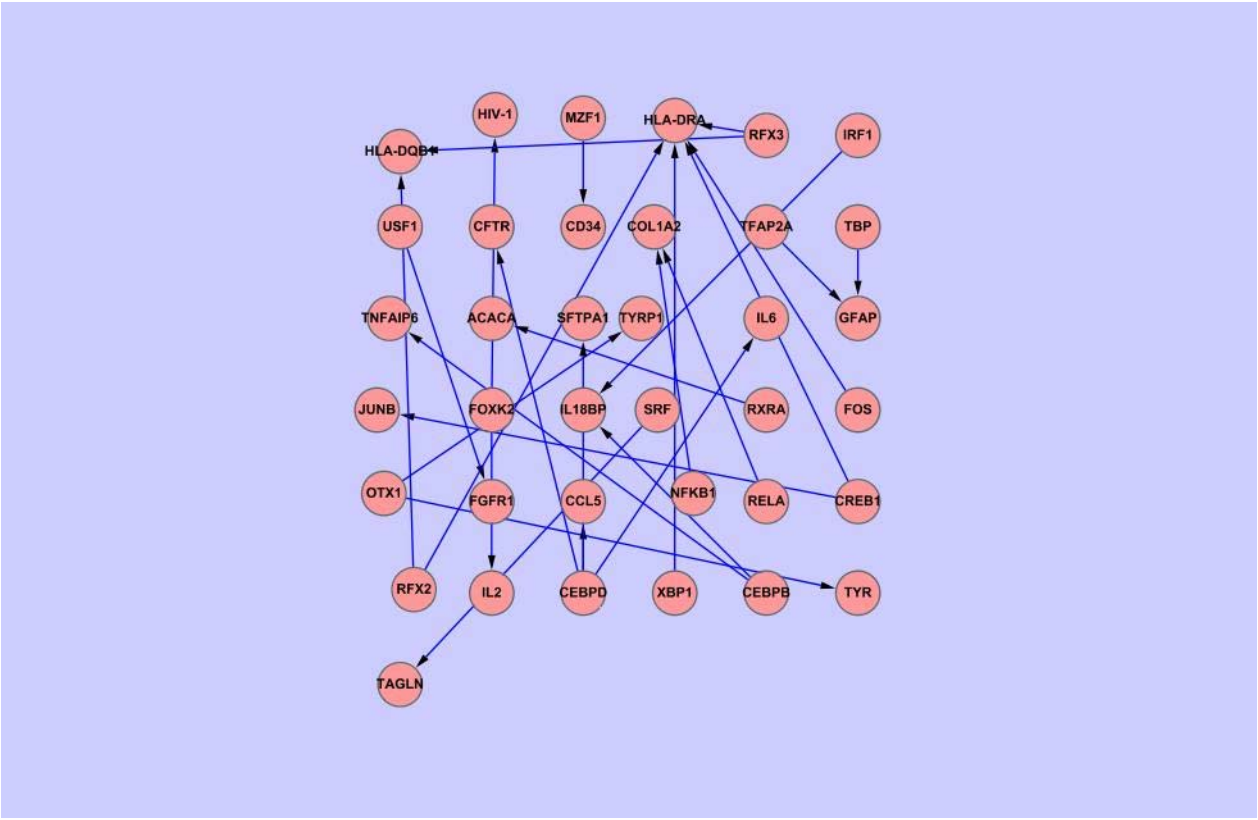


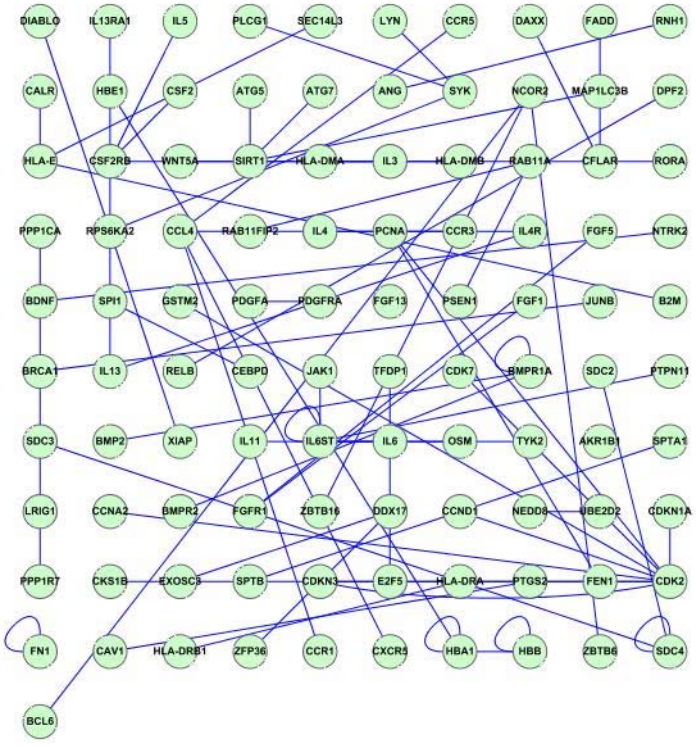
Type:Brain and CNS Cancer Subtype:Diffuse Astrocytoma



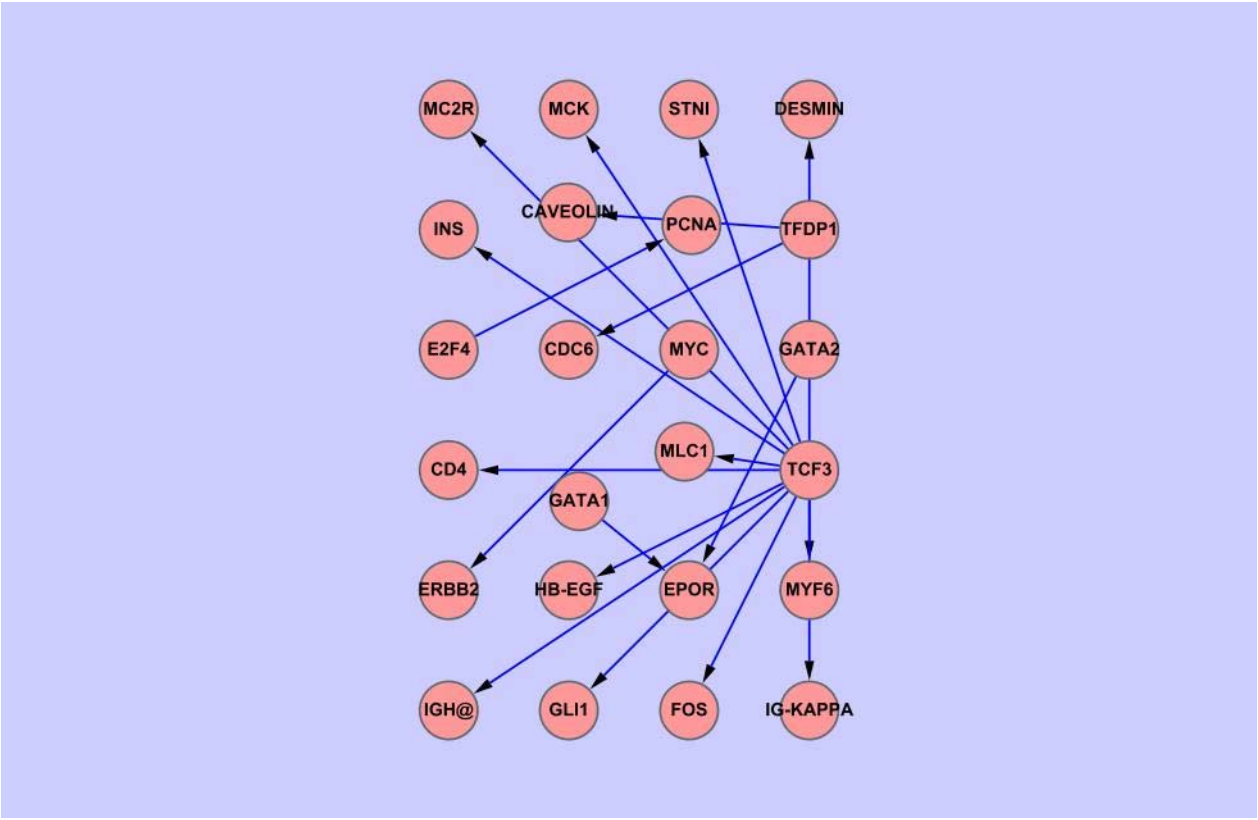


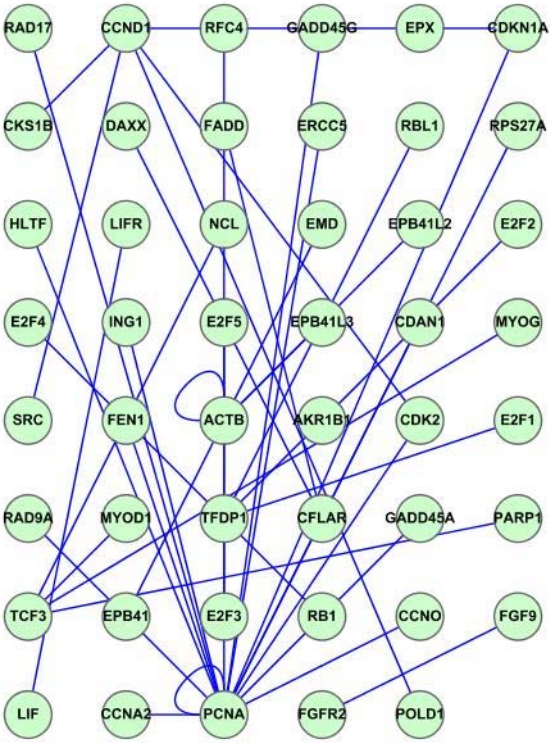
Type:Brain and CNS Cancer Subtype:Glioblastoma



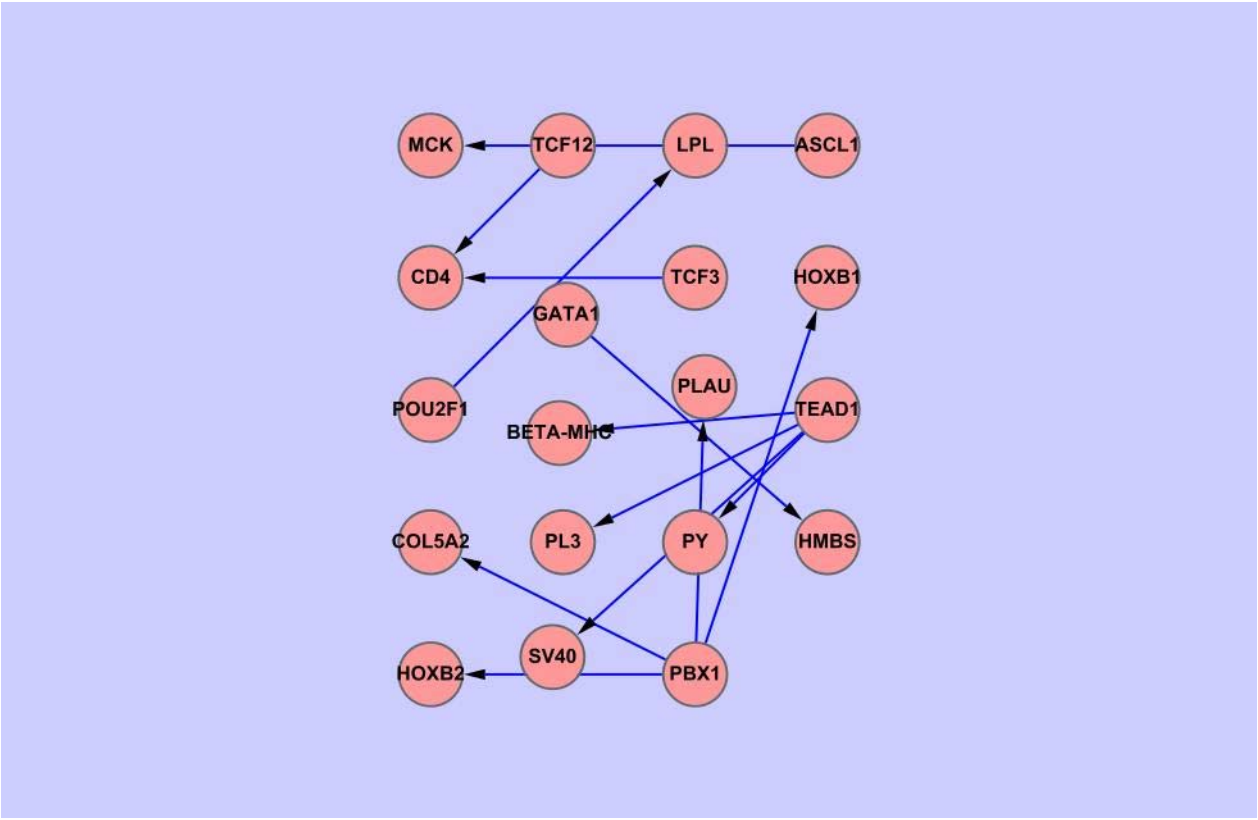


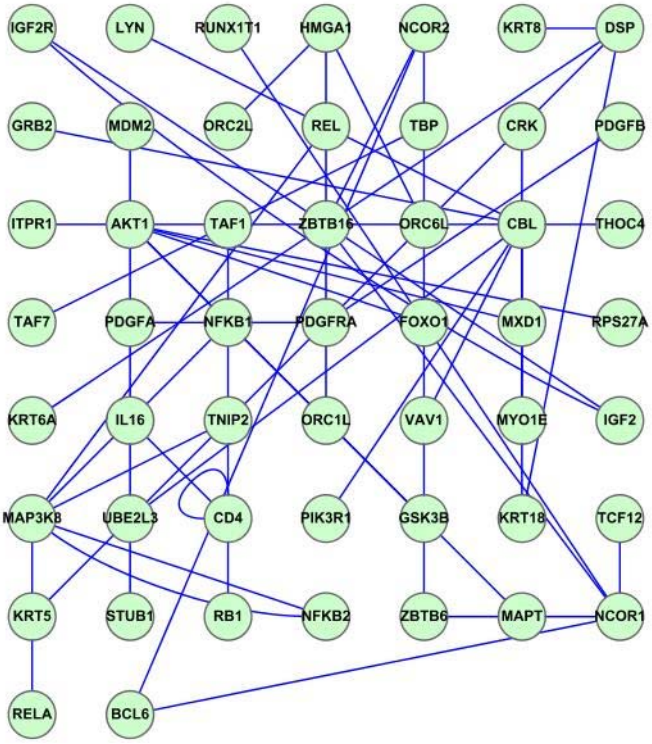
Type:Brain and CNS Cancer Subtype:Meningioma



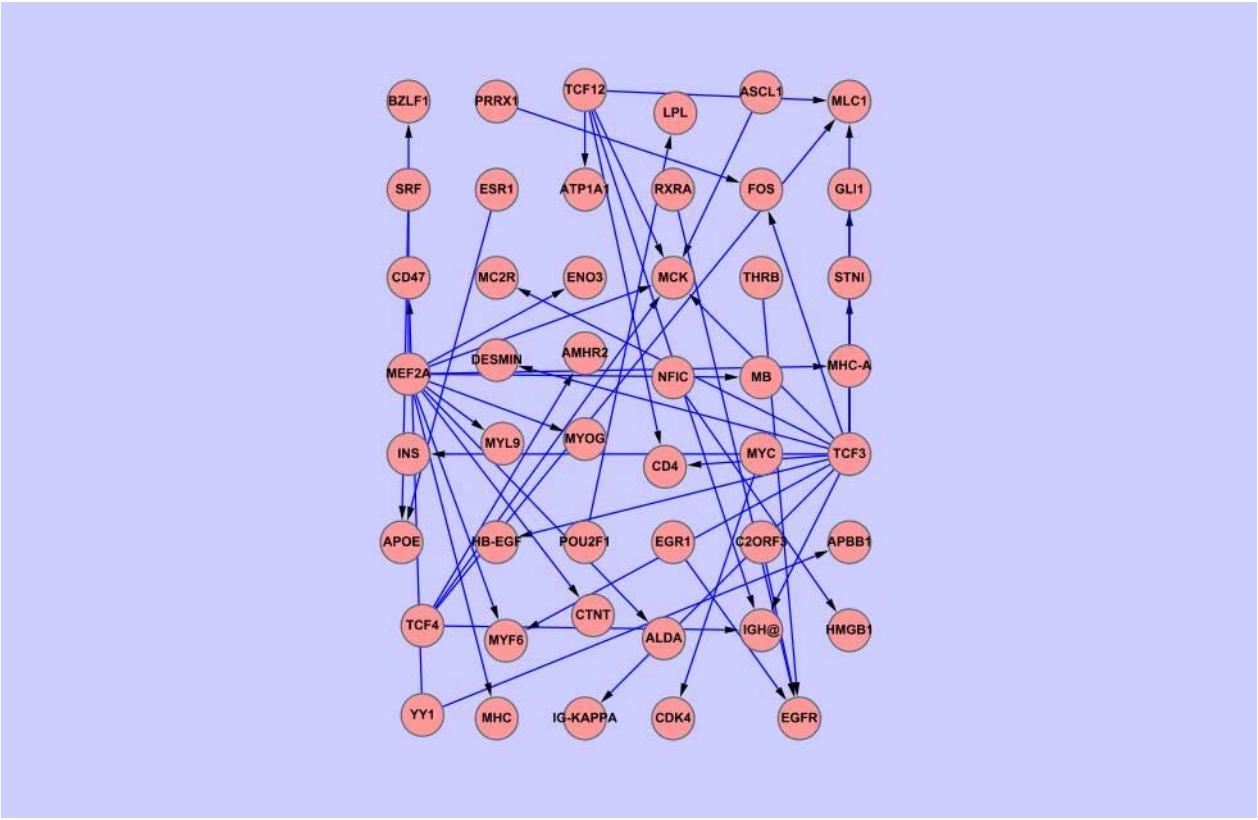


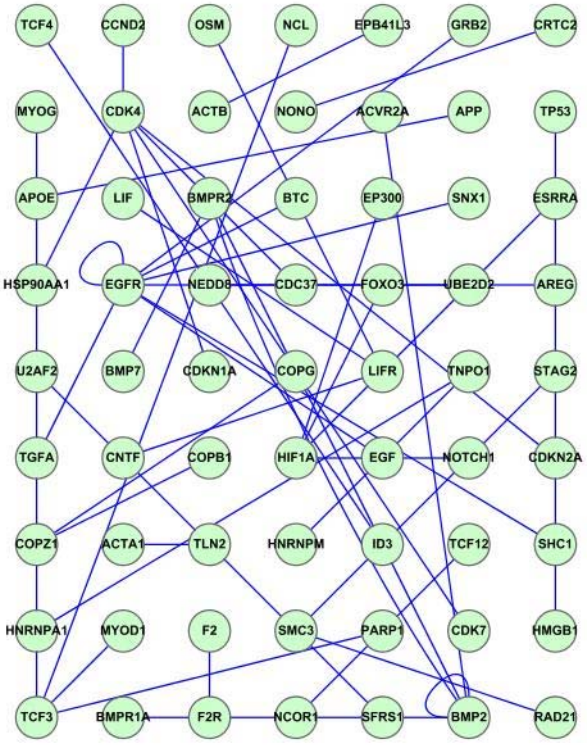
Type:Brain and CNS Cancer Subtype:Oligoastrocytoma



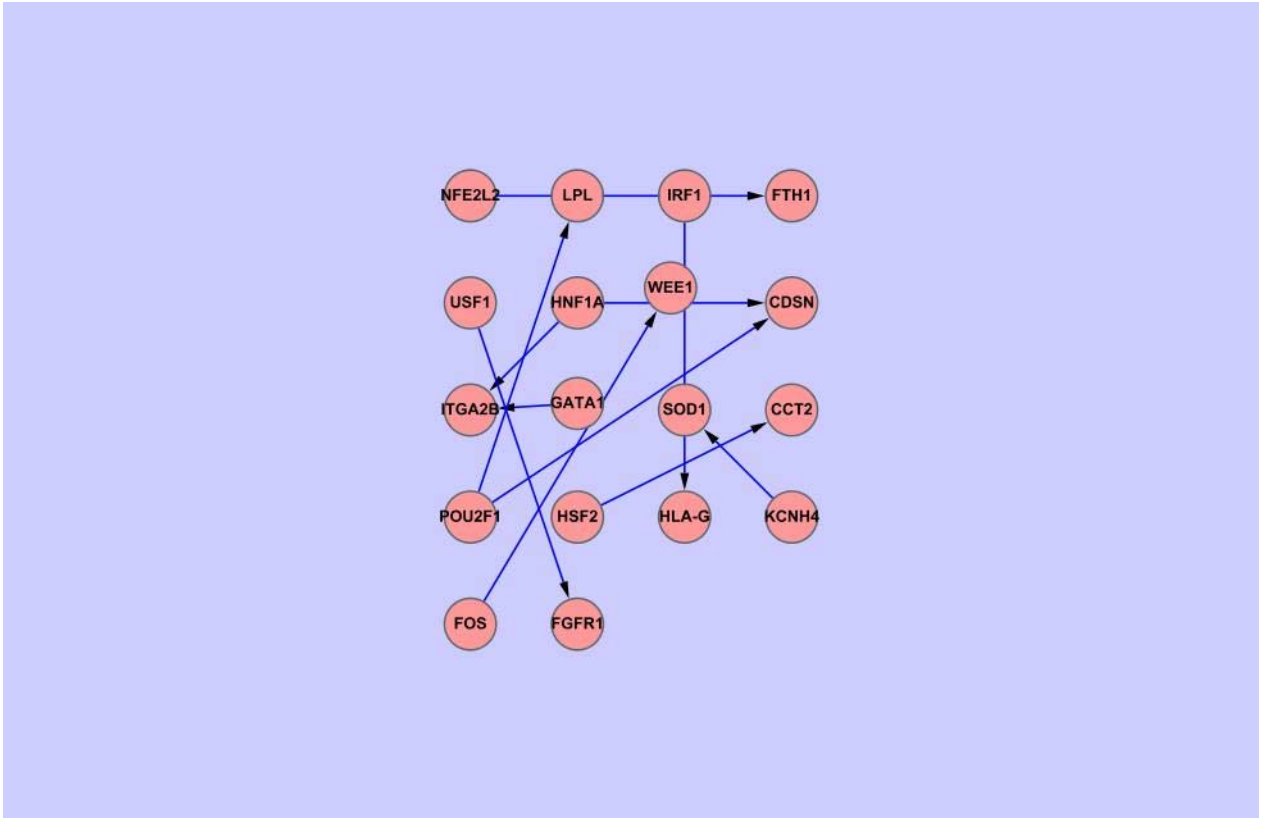


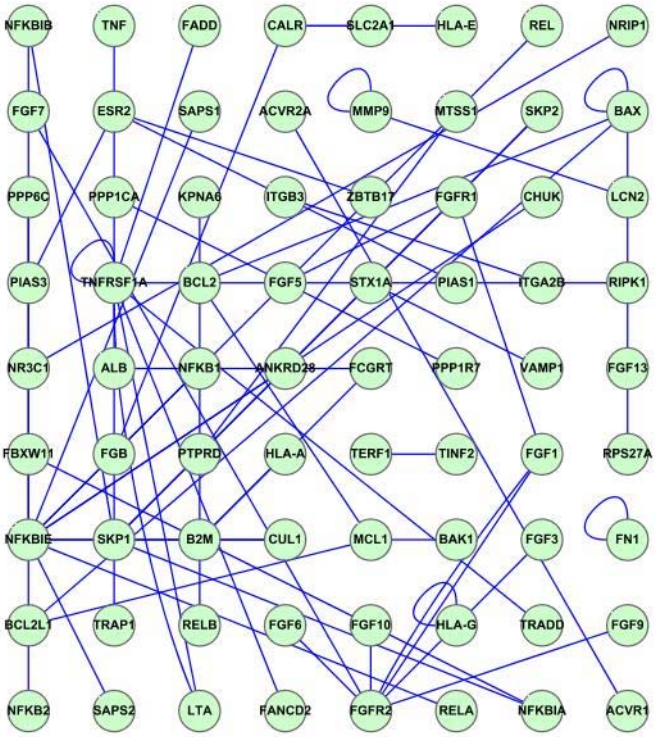
Type:Brain and CNS Cancer Subtype:Oligodendroglioma



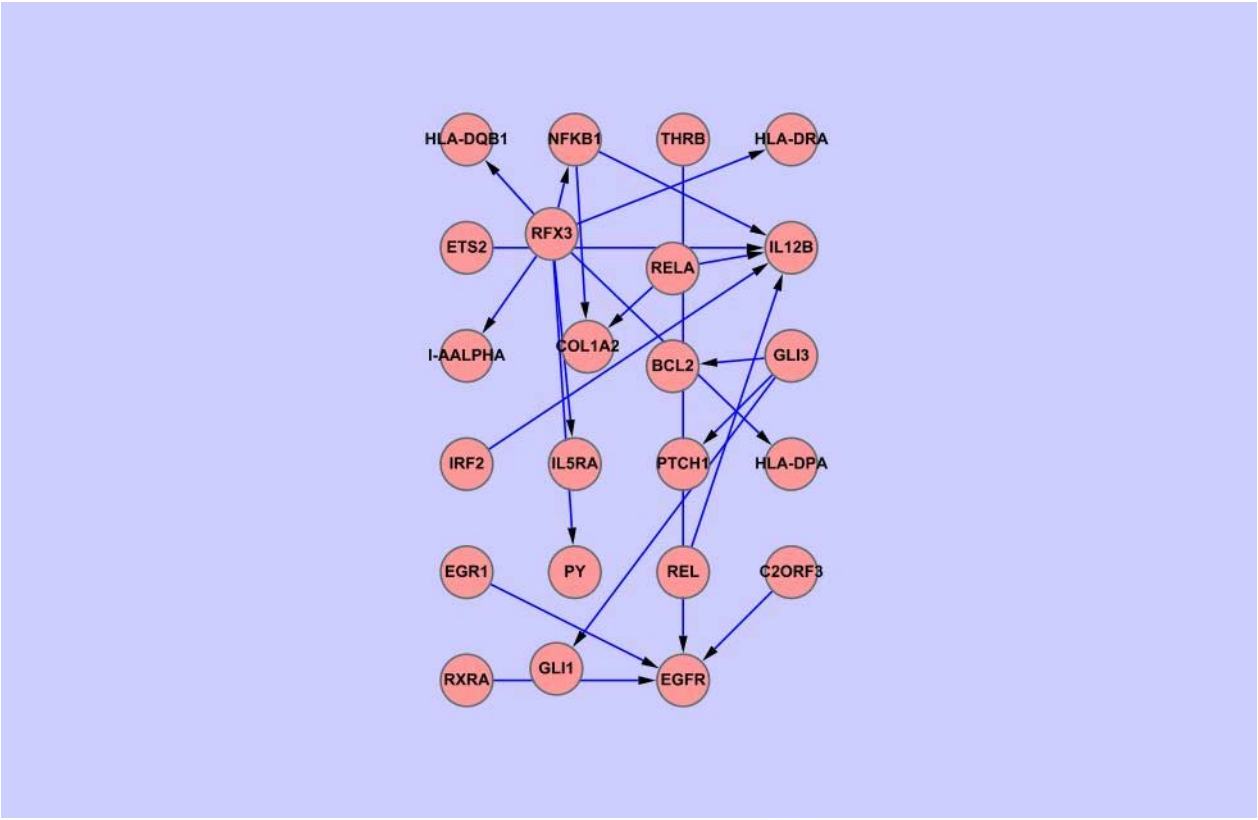


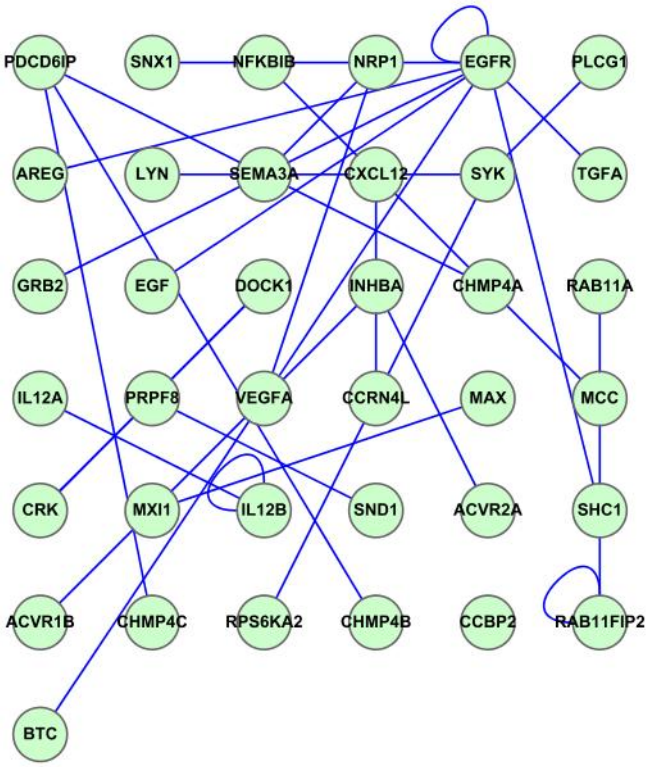
Type:Brain and CNS Cancer Subtype:Pilocytic Astrocytoma



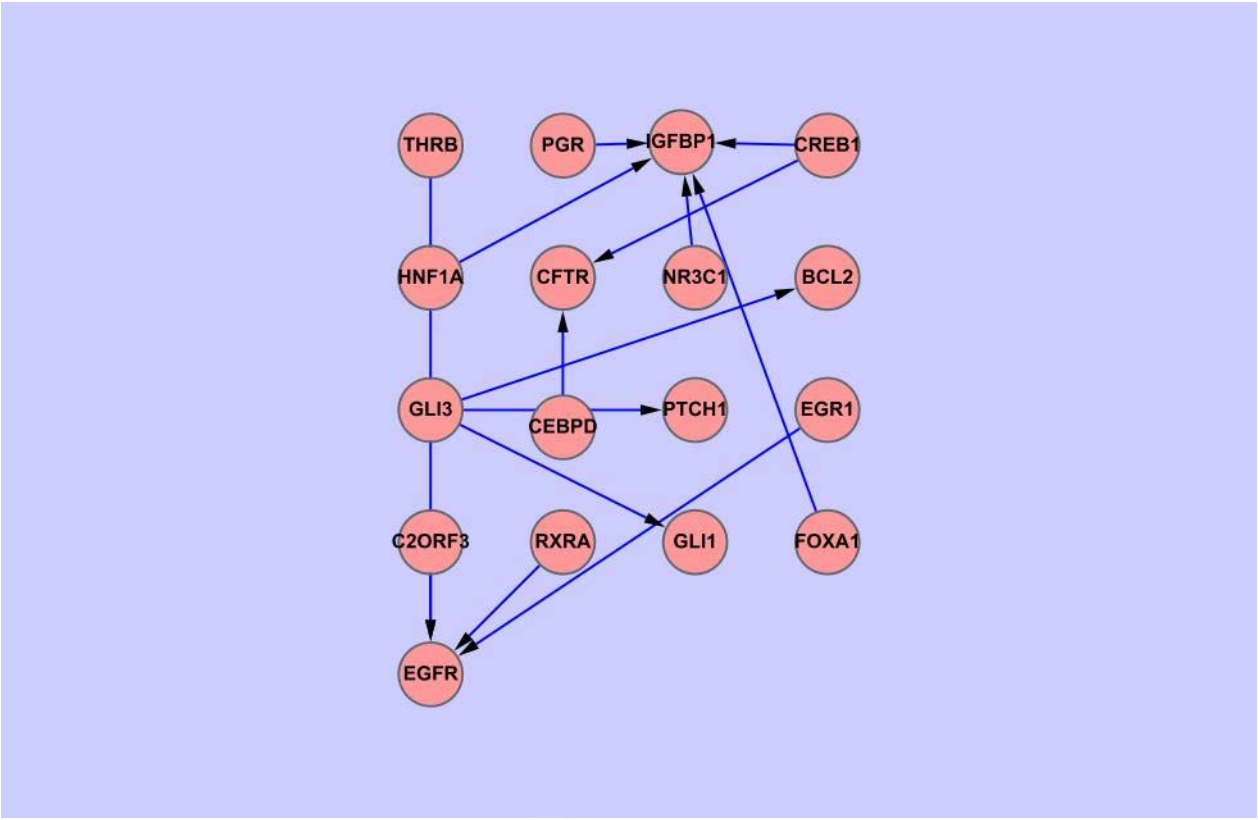


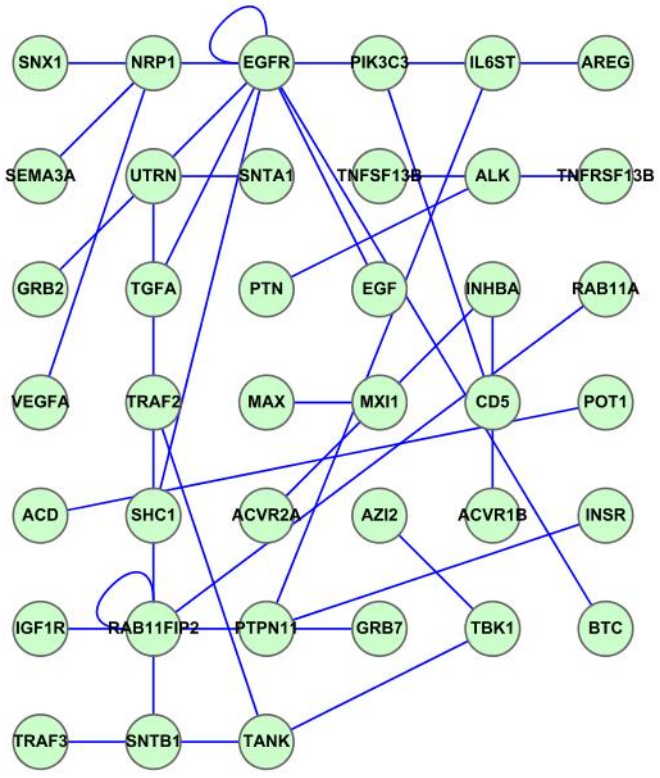
Type:Brain and CNS Cancer Subtype:Primary Glioblastoma



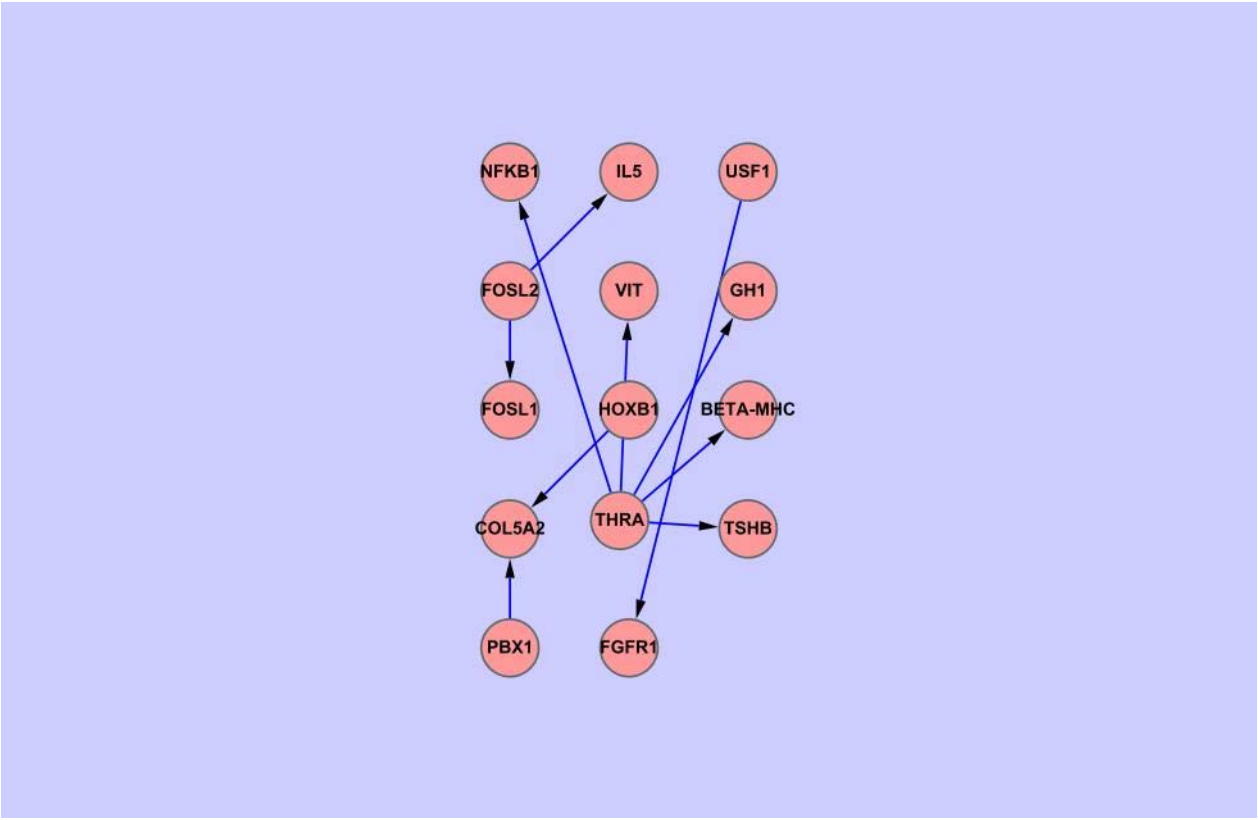


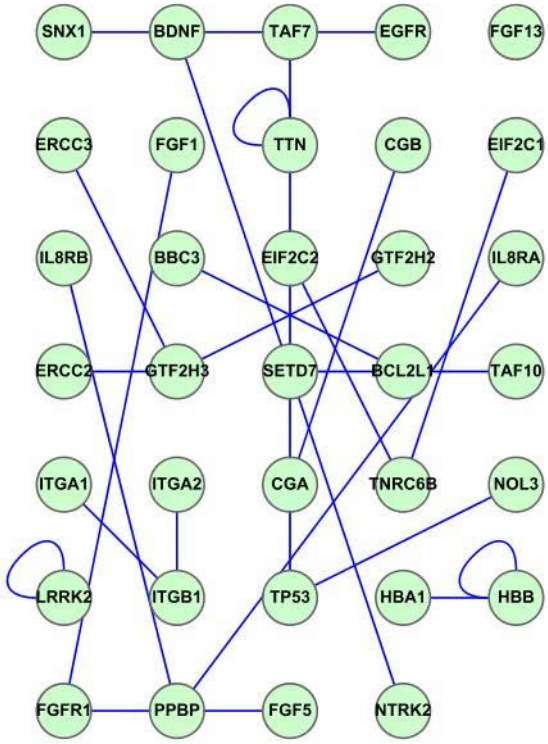
Type:Brain and CNS Cancer Subtype:Secondary Glioblastoma



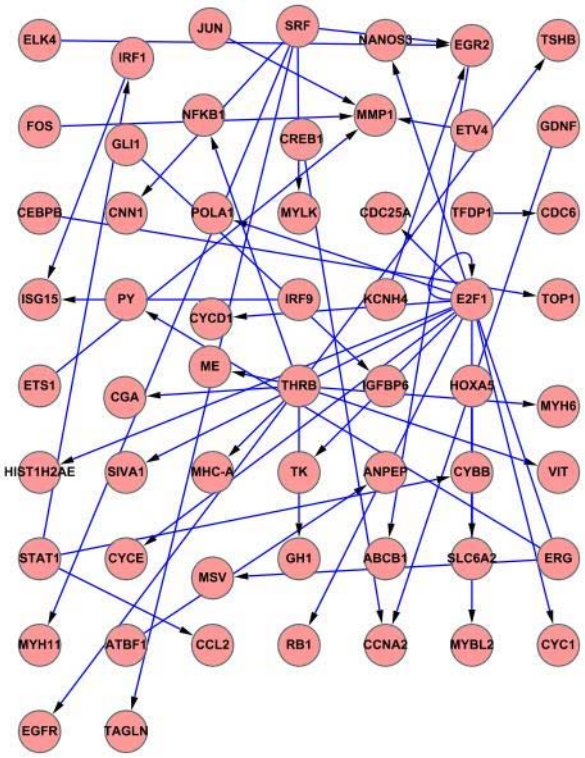


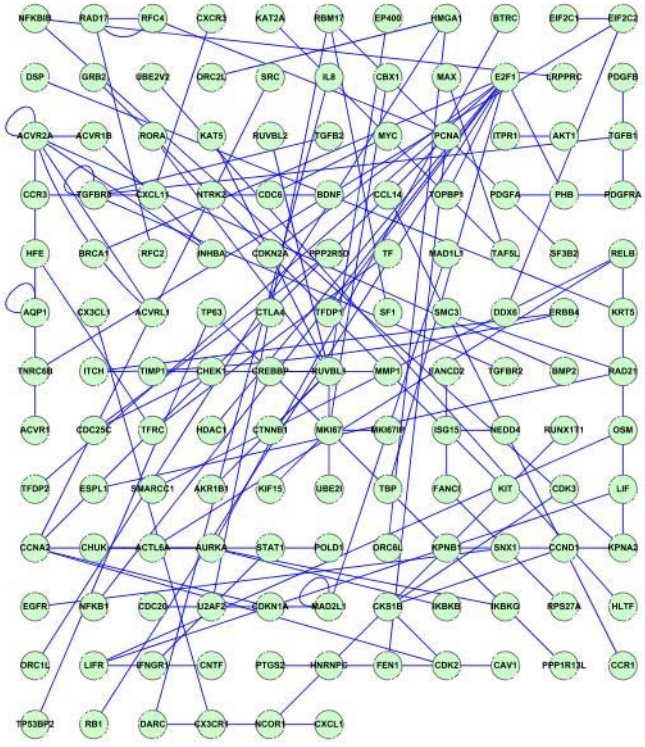
Type:Breast Cancer Subtype:Ductal Breast Carcinoma in Situ



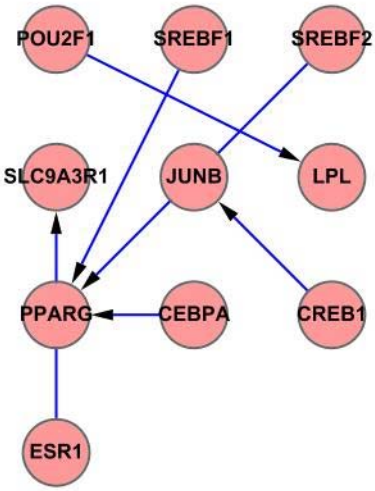


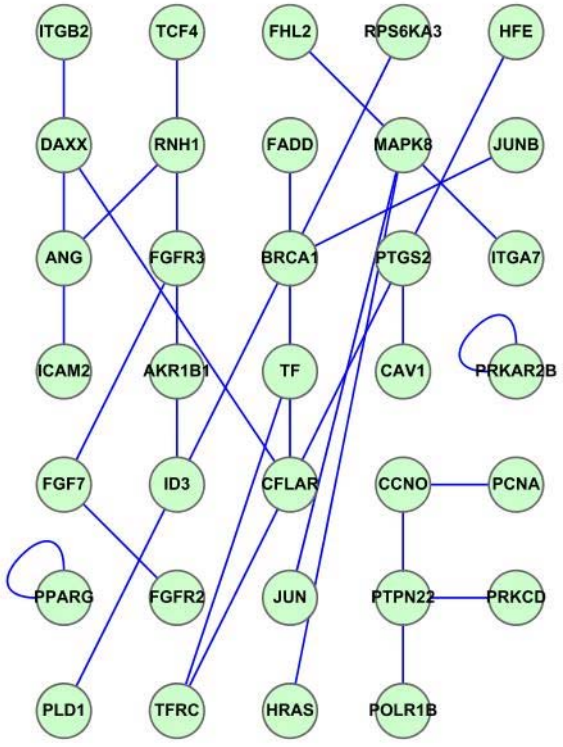
Type:Breast Cancer Subtype:Ductal Breast Carcinoma



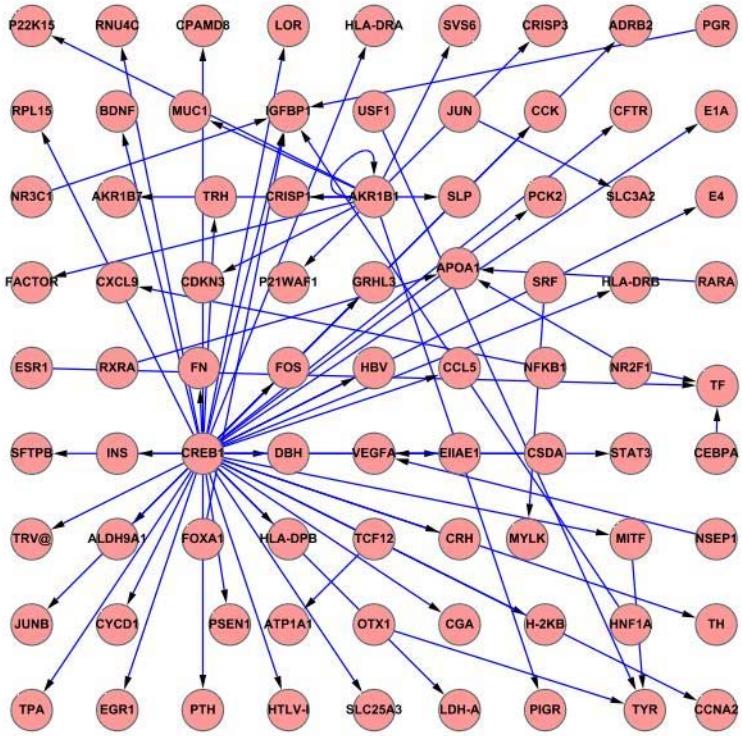


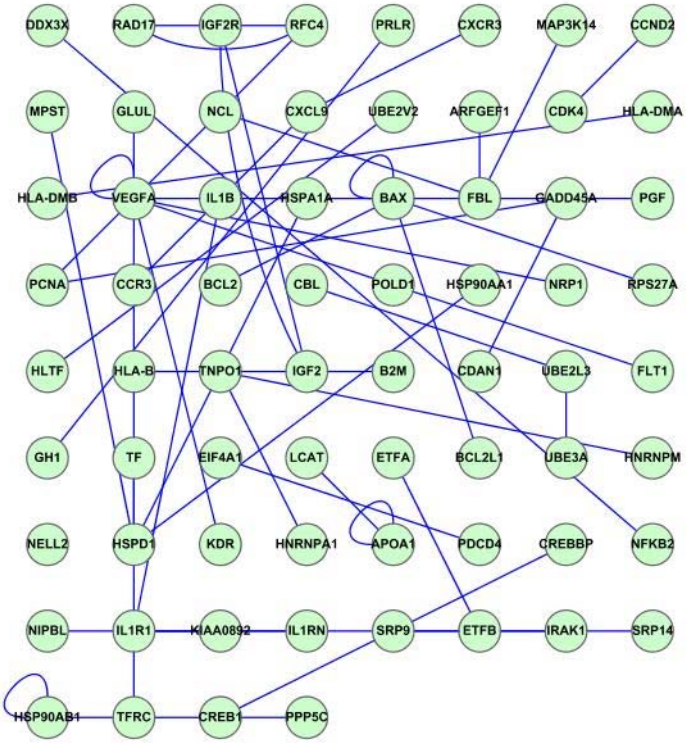
Type:Breast Cancer Subtype:Fibroadenoma



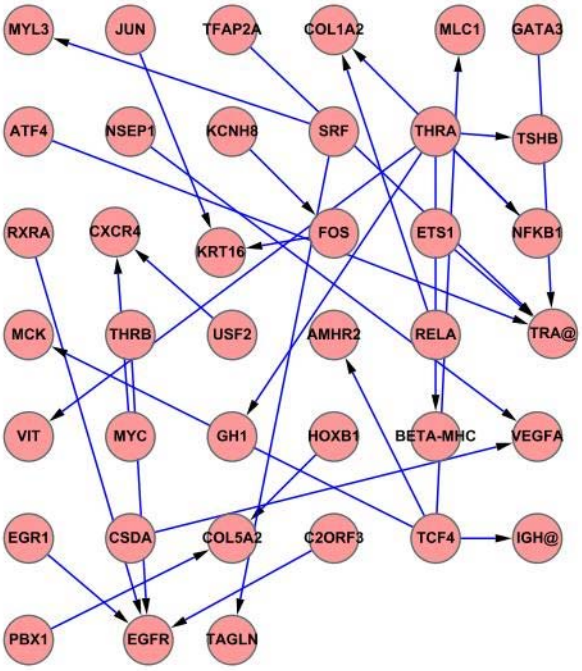


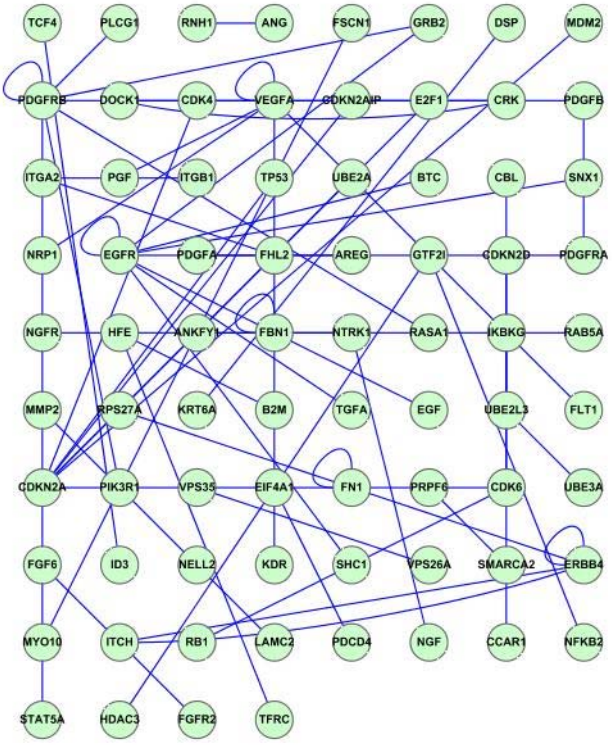
Type:Breast Cancer Subtype:Invasive Breast Carcinoma



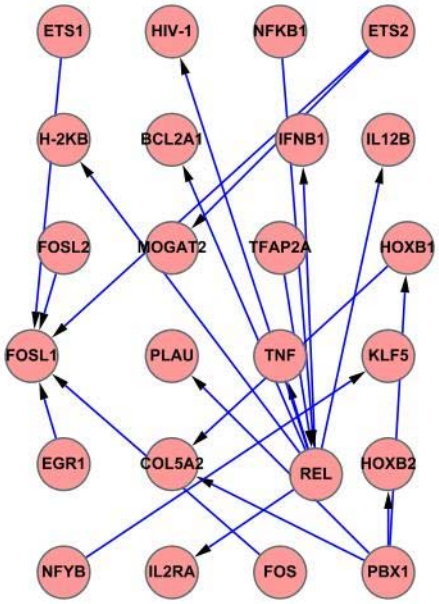


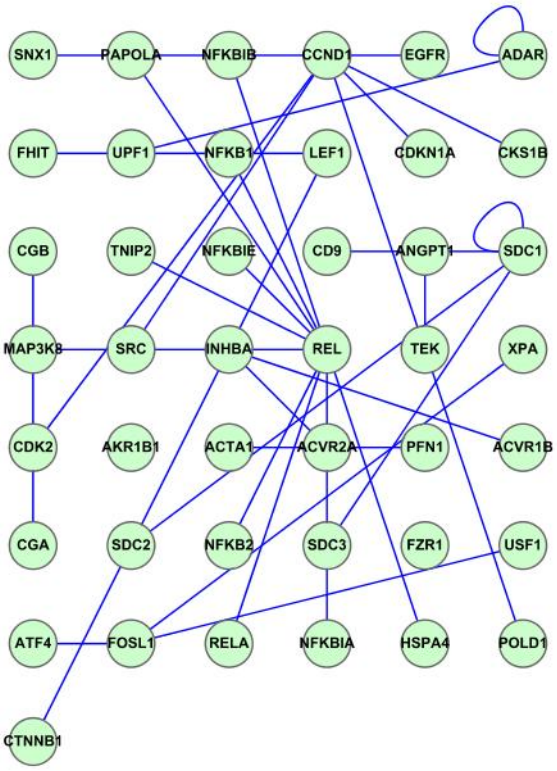
Type:Breast Cancer Subtype:Invasive Lobular Breast Carcinoma



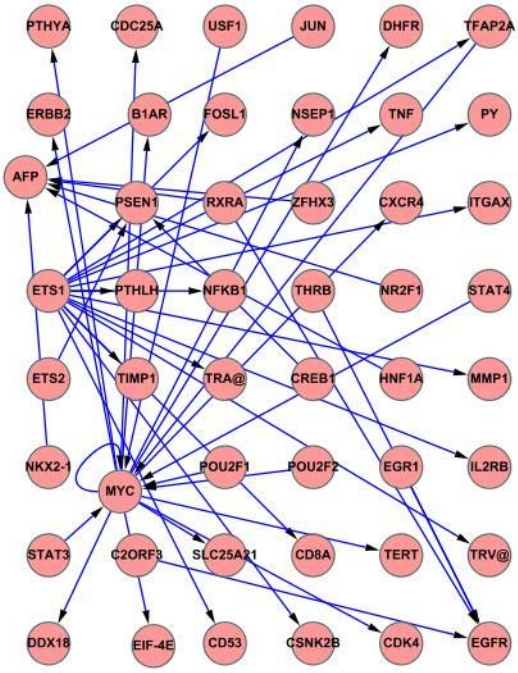


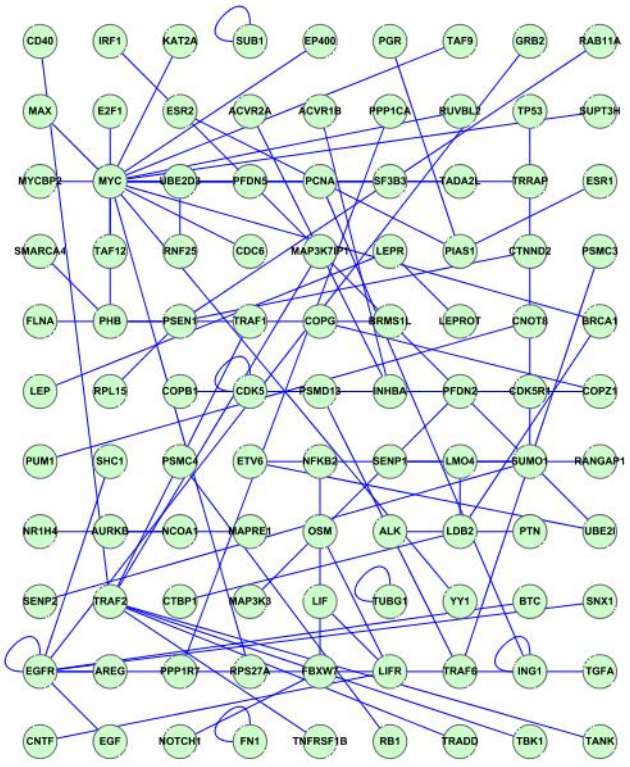
Type:Breast Cancer Subtype:Invasive Mixed Breast Carcinoma



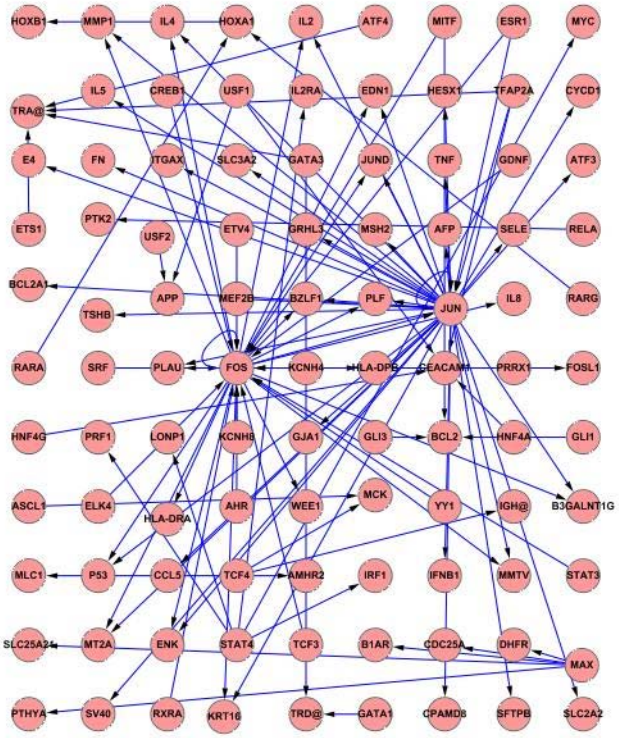


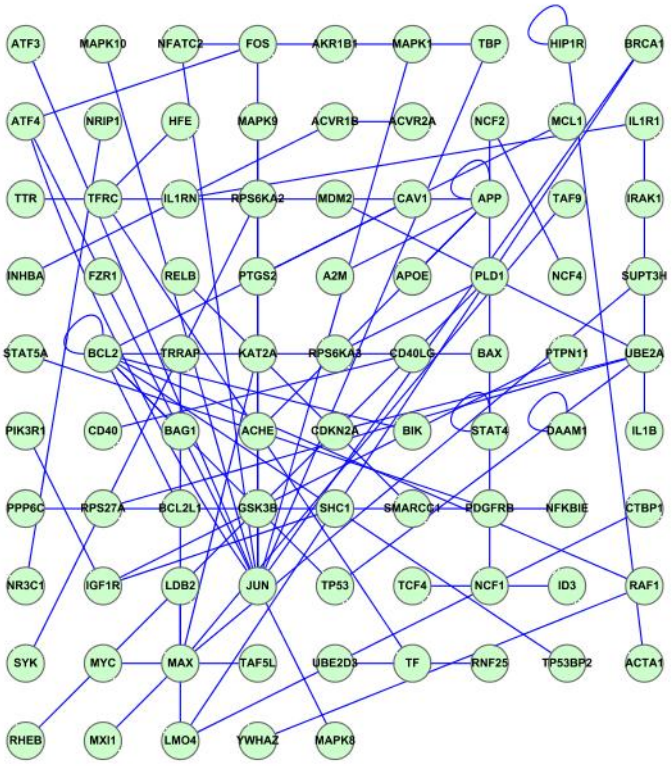
Type:Breast Cancer Subtype:Lobular Breast Carcinoma



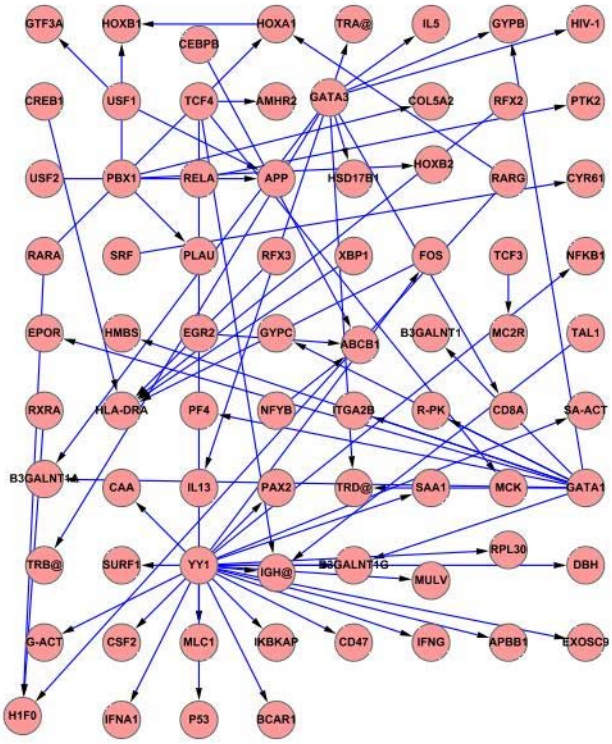


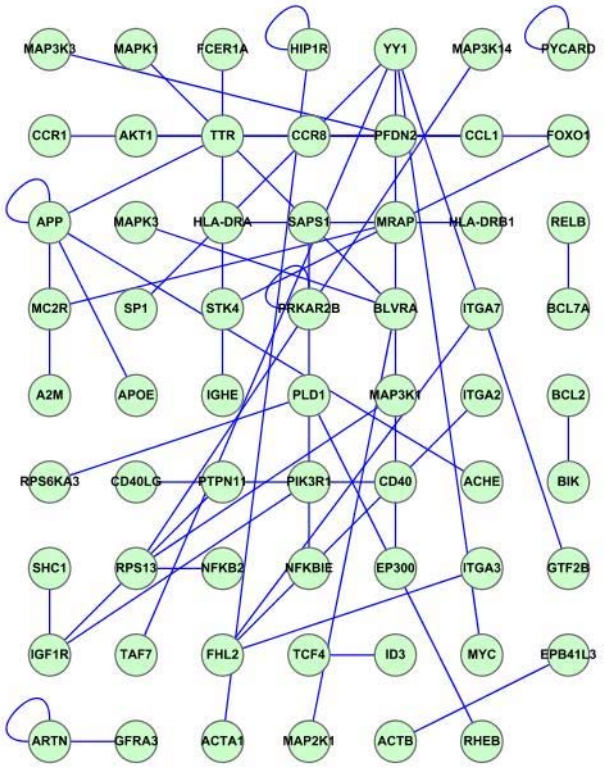
Type:Leukemia Subtype:Acute Adult T-Cell Leukemia/Lymphoma



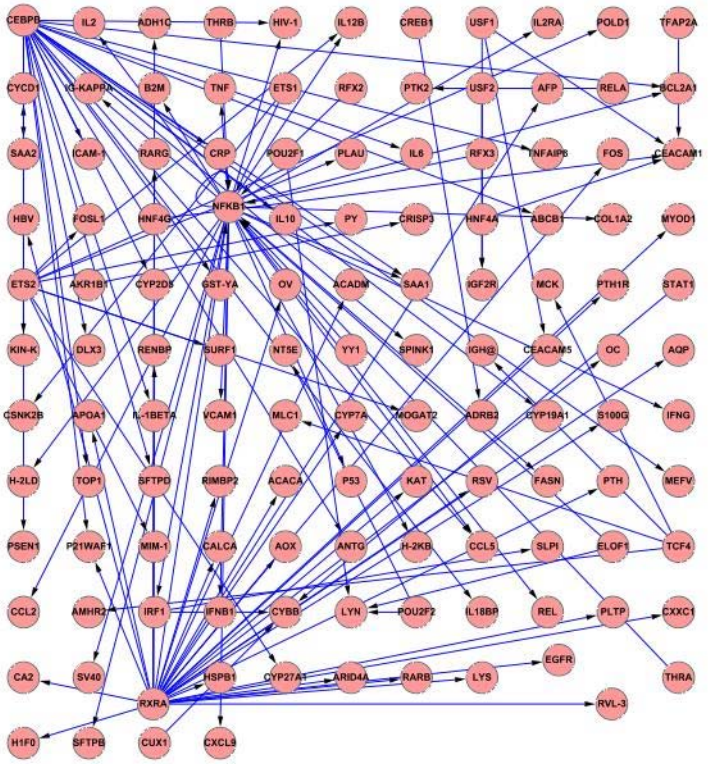


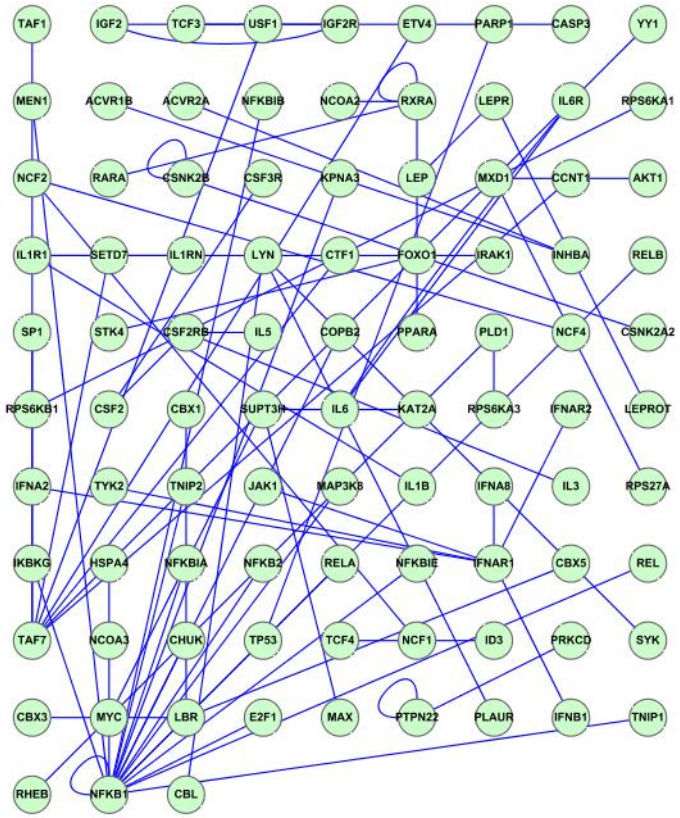
Type:Leukemia Subtype:Acute Myeloid Leukemia



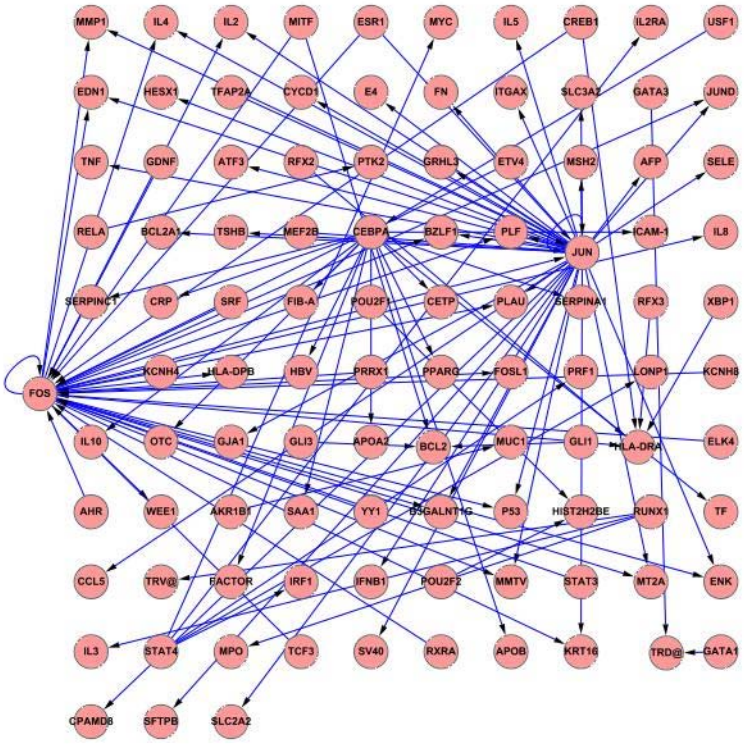


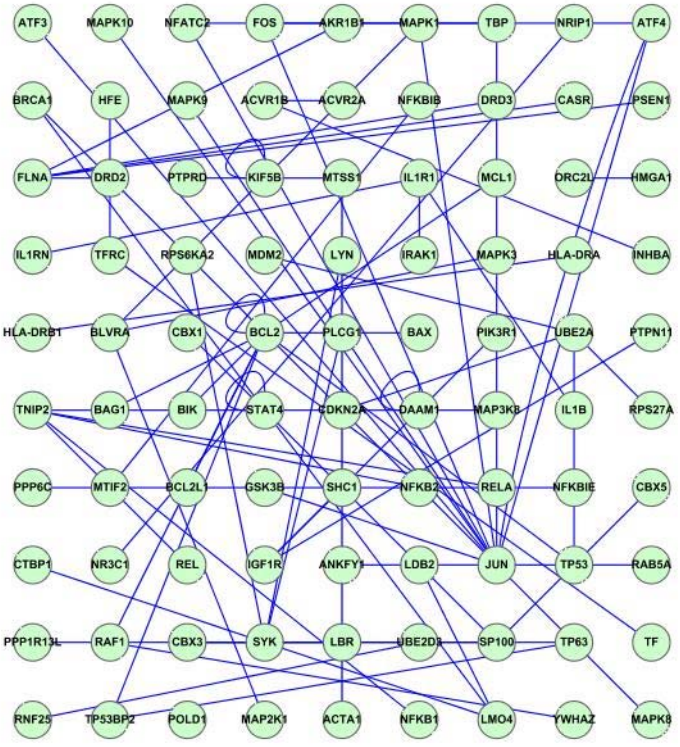
Type:Leukemia Subtype:B-Cell Acute Lymphoblastic Leukemia



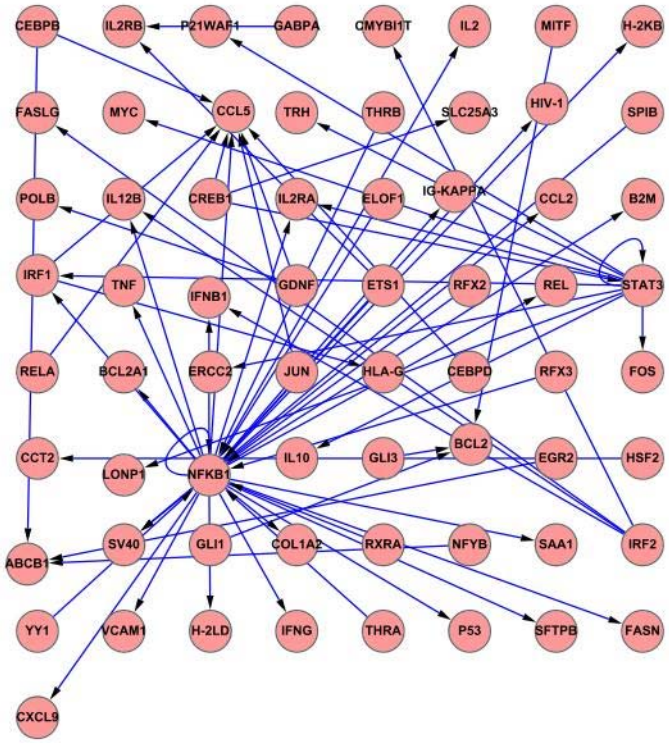


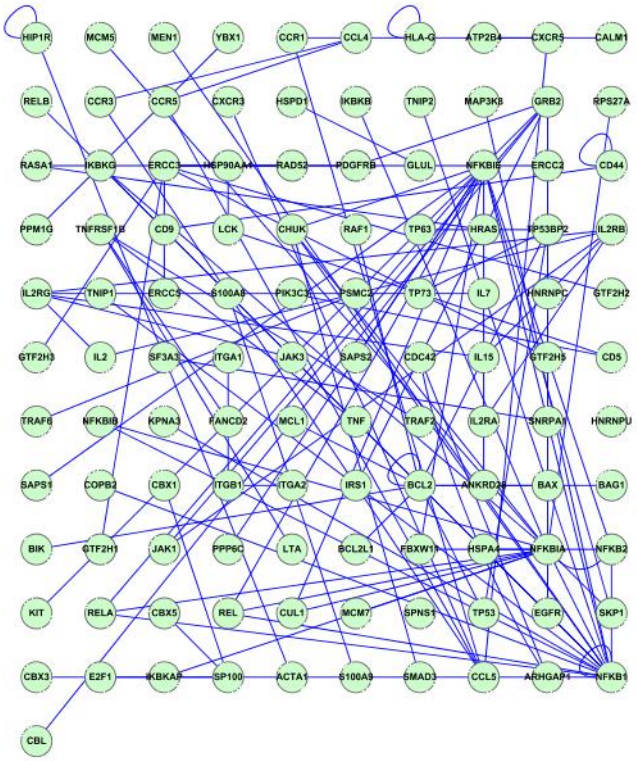
Type:Leukemia Subtype:Chronic Adult T-Cell Leukemia/Lymphoma



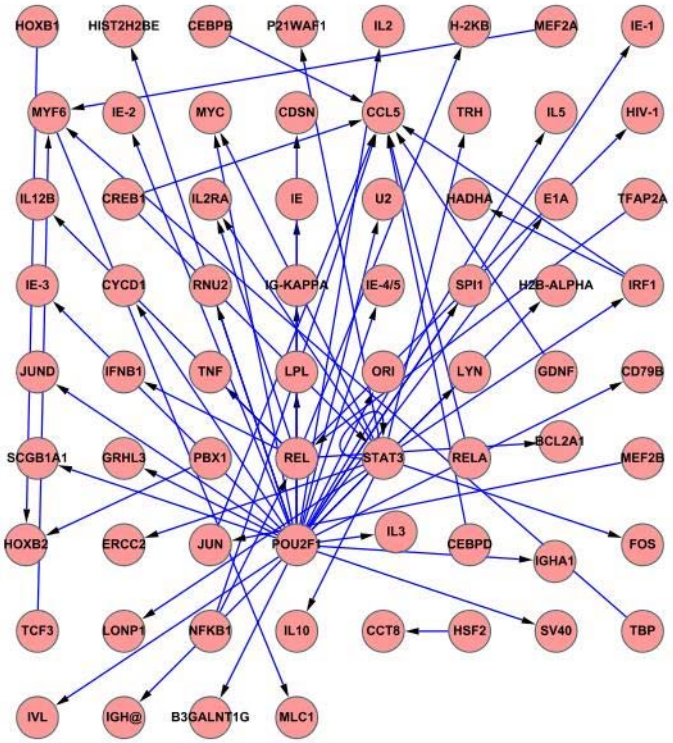


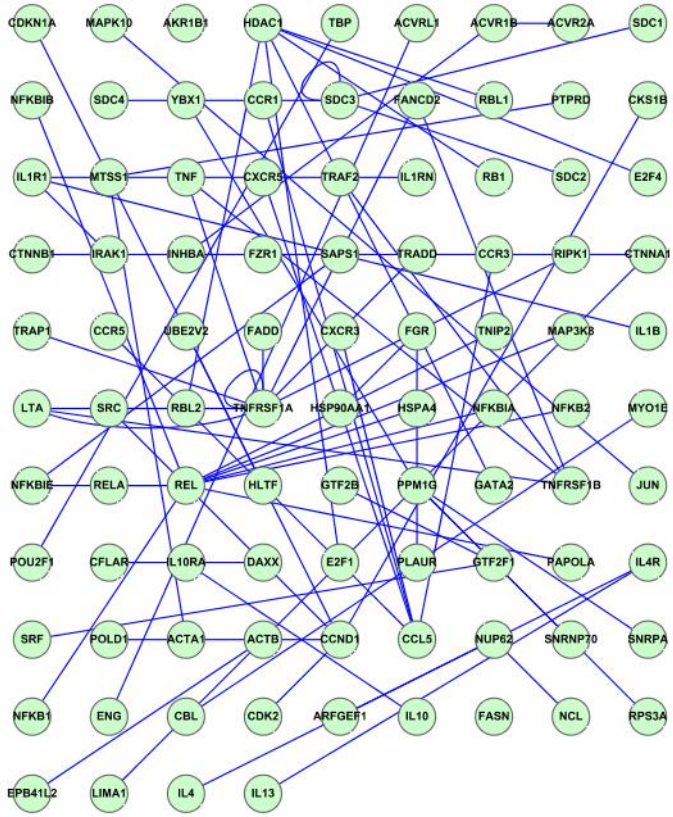
Type:Leukemia Subtype:Chronic Lymphocytic Leukemia



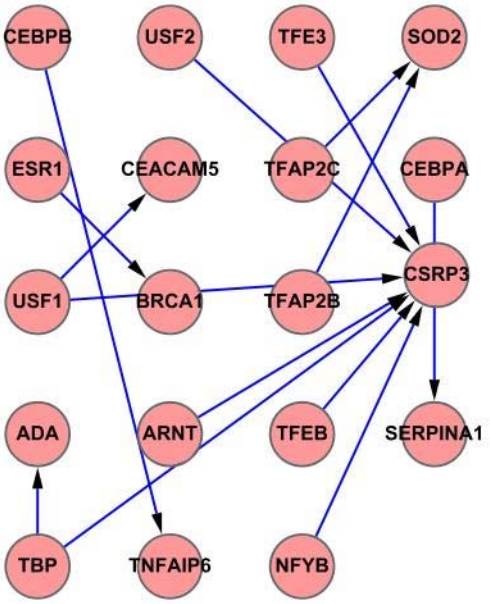


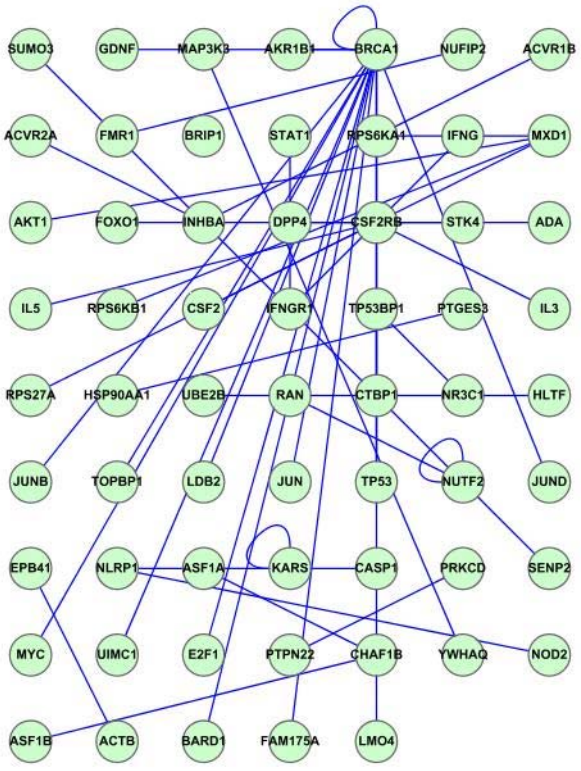
Type:Leukemia Subtype:Hairy Cell Leukemia



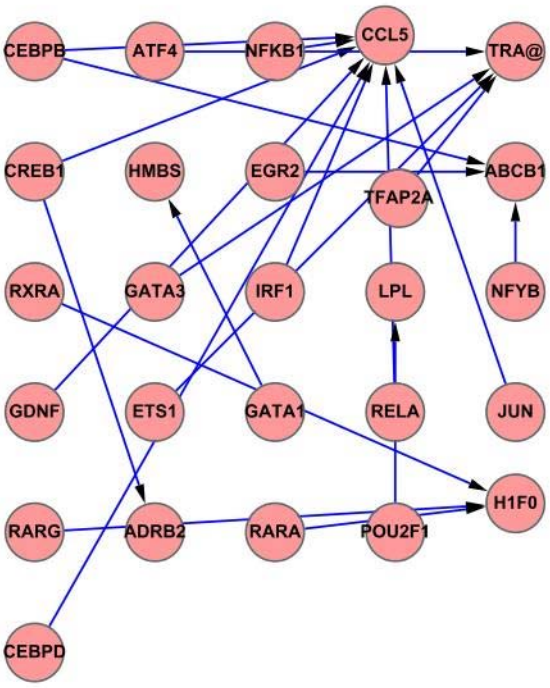


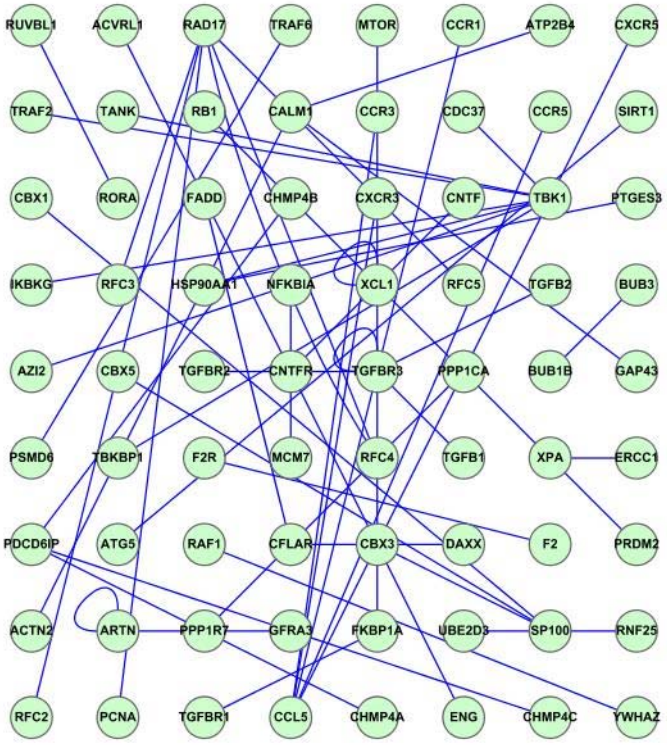
Type:Leukemia Subtype:T-Cell Acute Lymphoblastic Leukemia



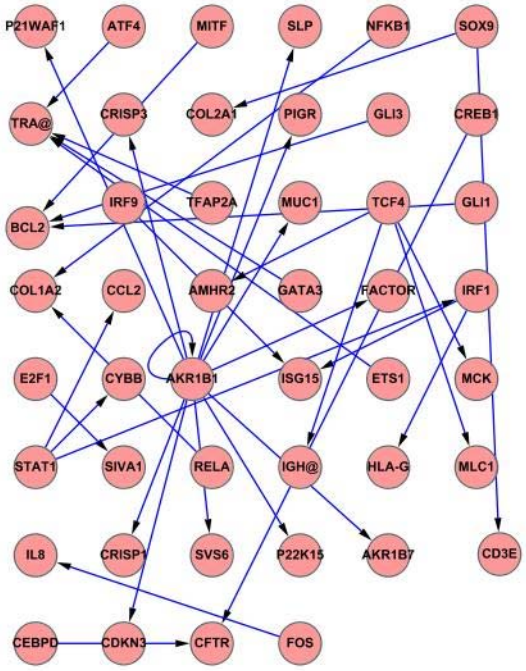


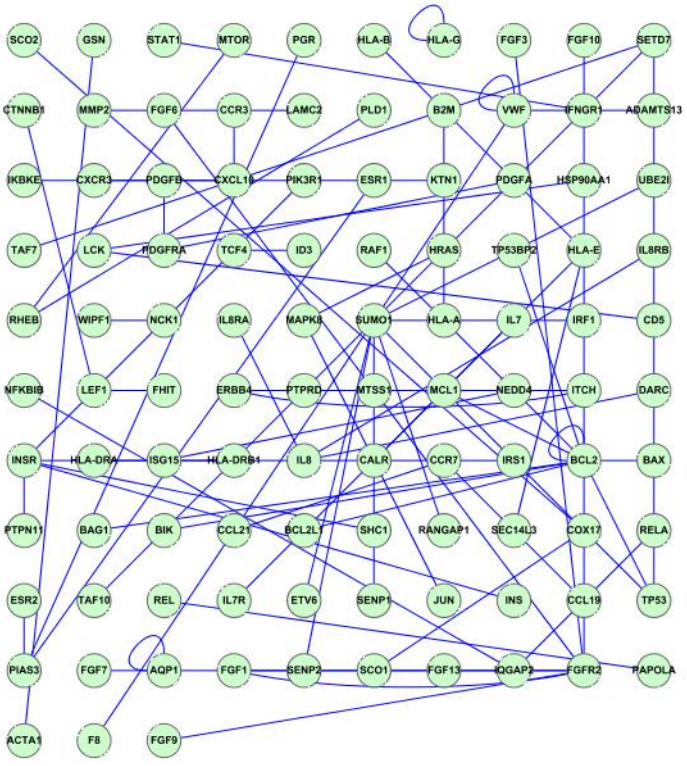
Type:Leukemia Subtype:T-Cell Prolymphocytic Leukemia



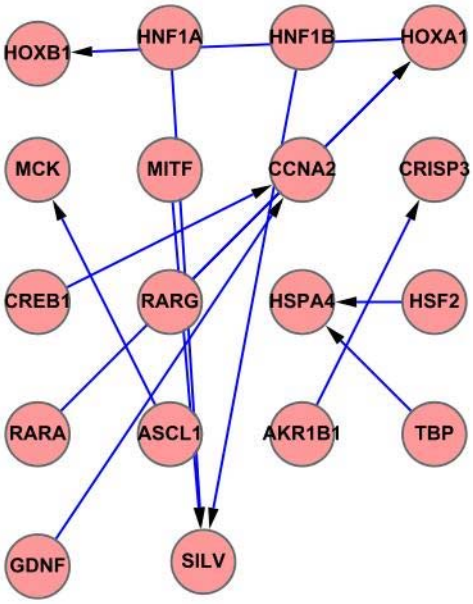


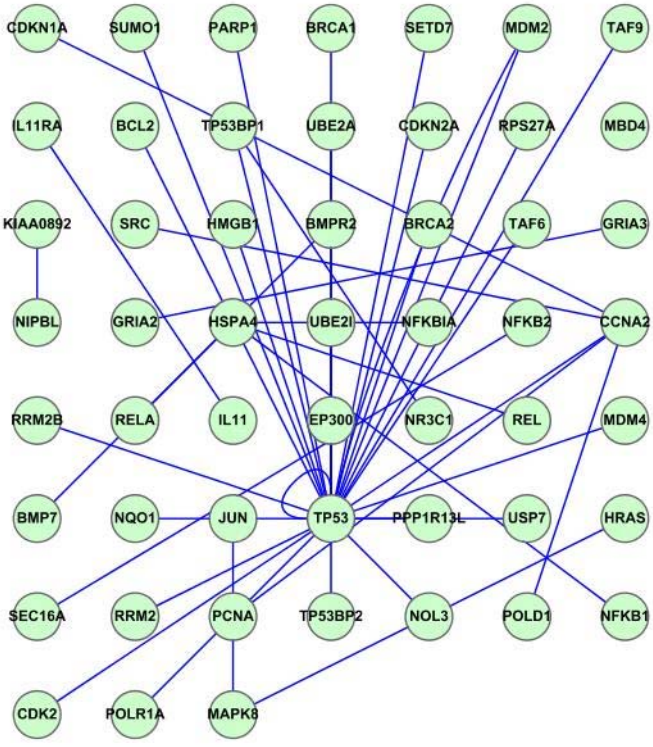
Type:Liver Cancer Subtype:Cirrhosis



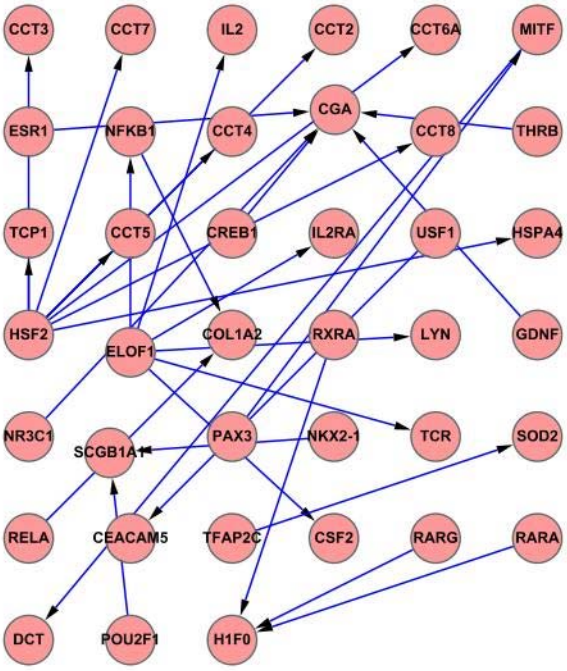


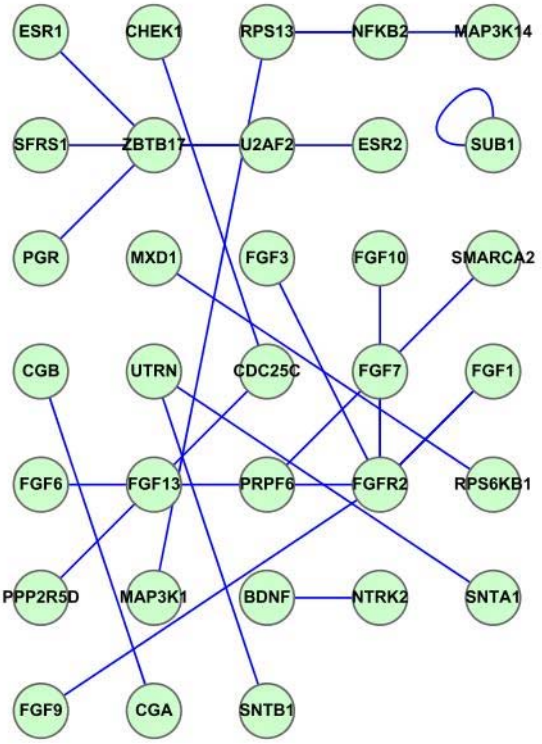
Type:Liver Cancer Subtype:Focal Nodular Hyperplasia of the Liver



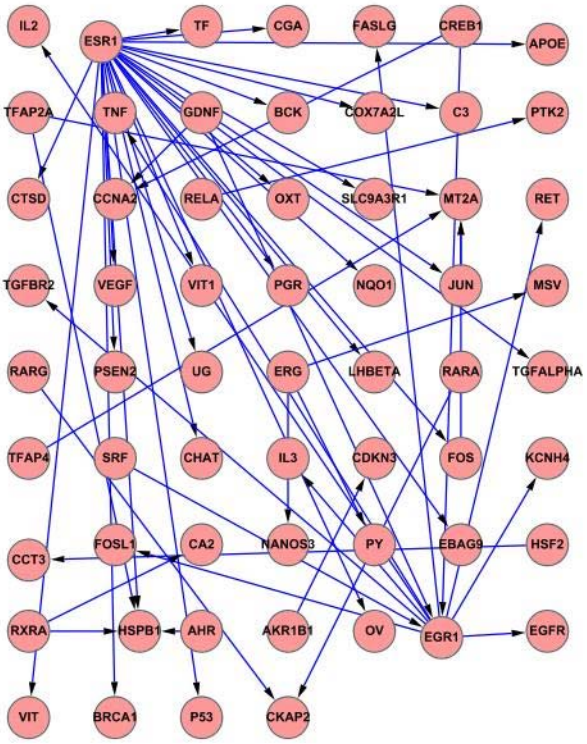


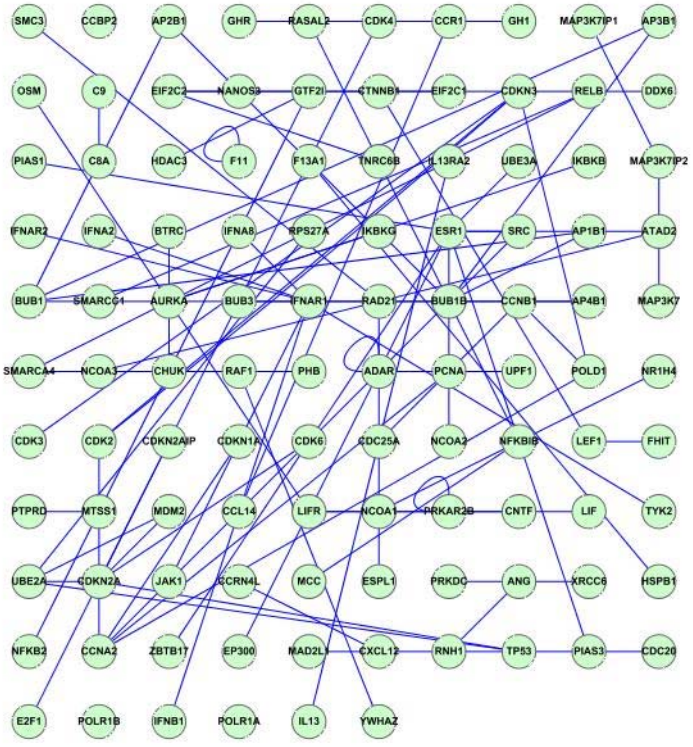
Type:Liver Cancer Subtype:Hepatocellular Adenoma



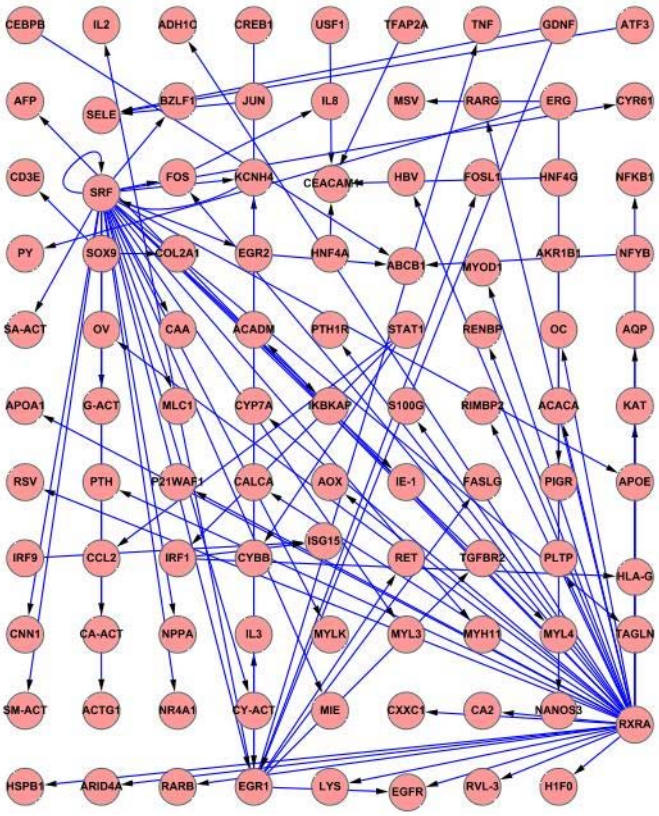


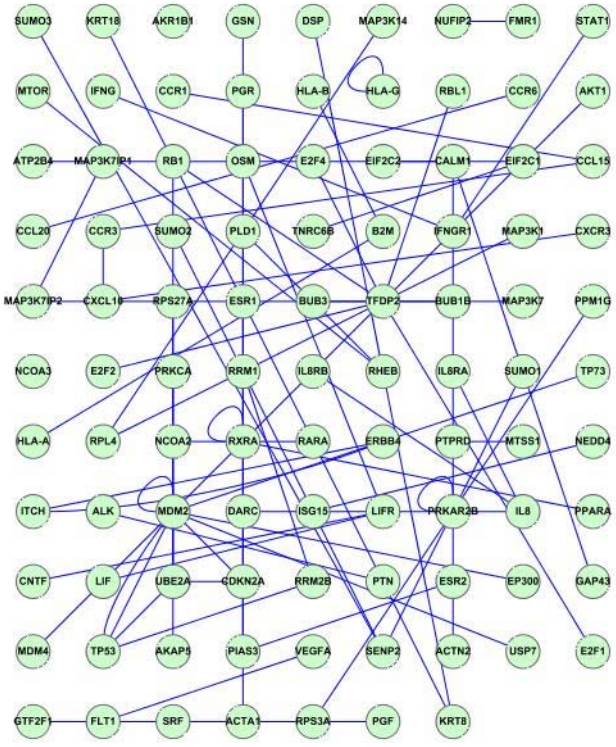
Type:Liver Cancer Subtype:Hepatocellular Carcinoma



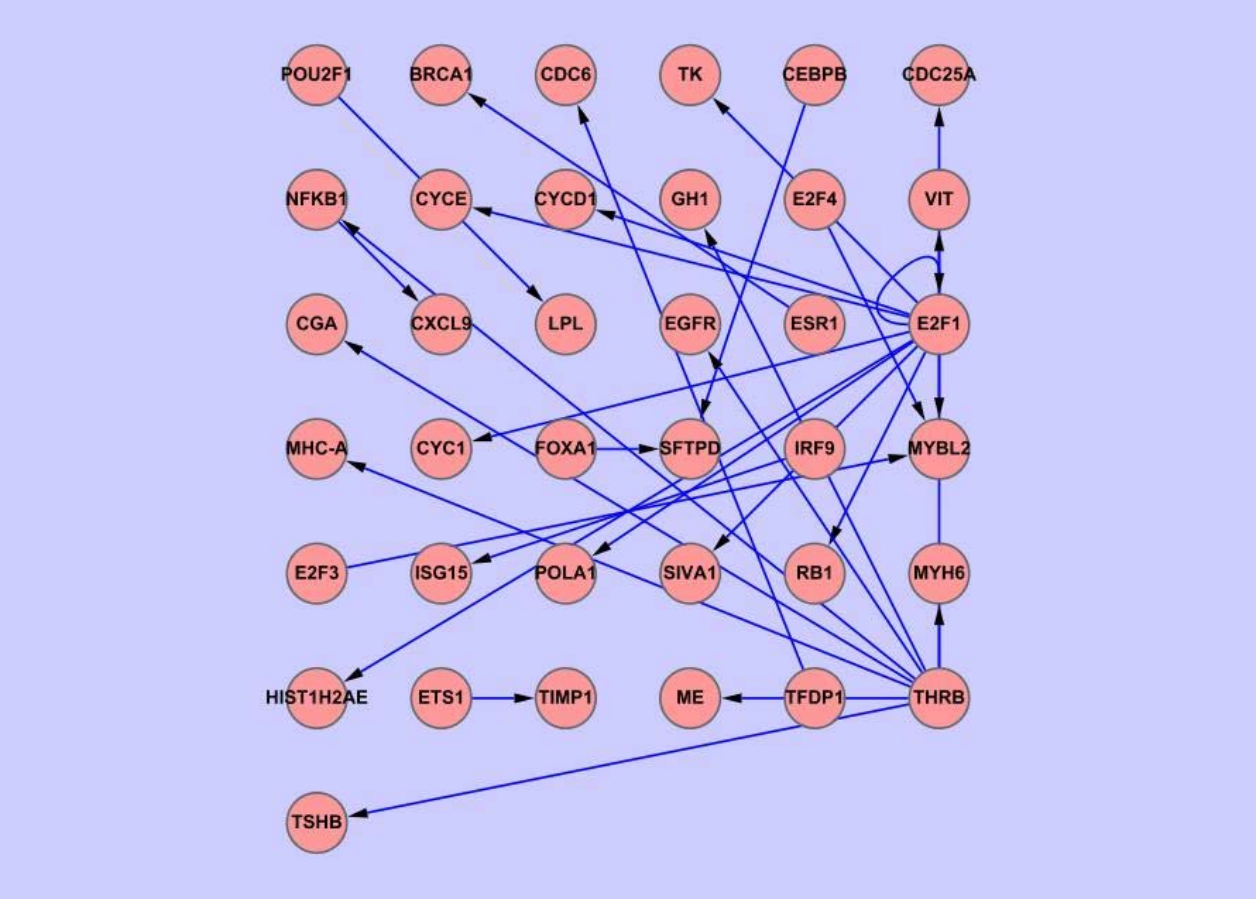


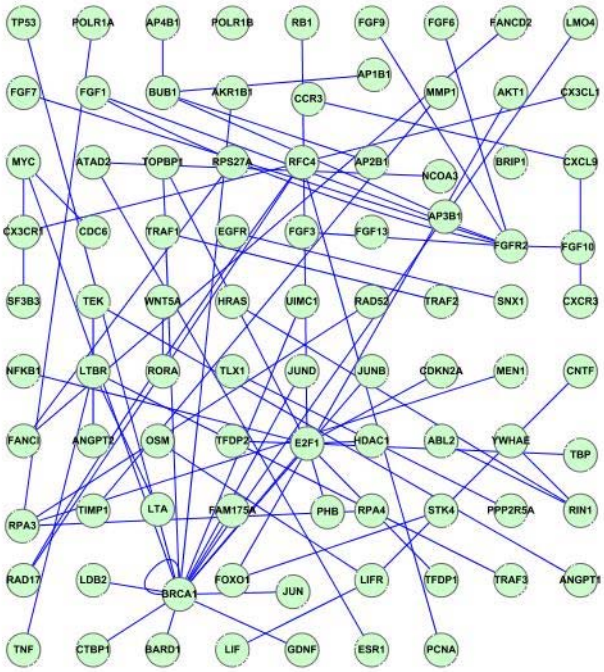
Type:Liver Cancer Subtype:Liver Cell Dysplasia



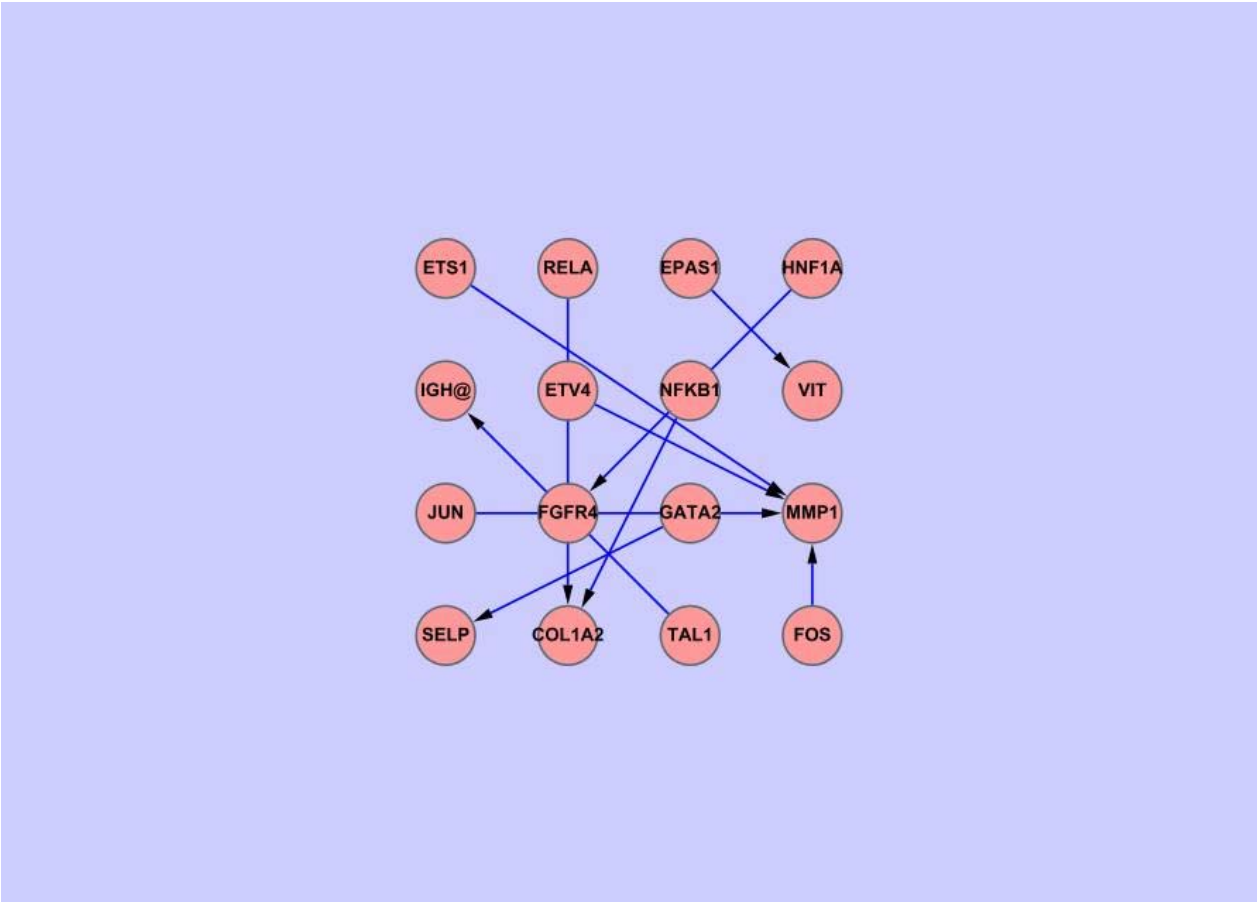


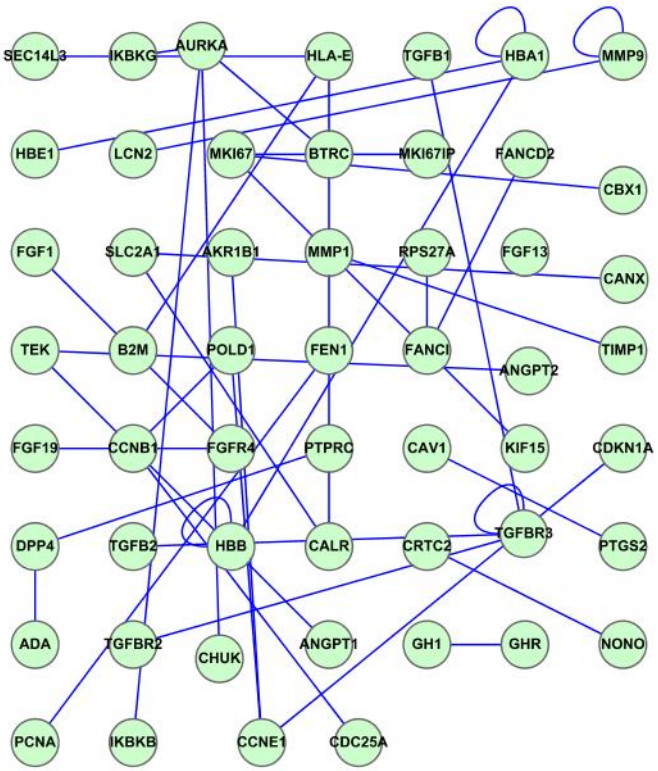
Type:Lung Cancer Subtype:Large Cell Lung Carcinoma



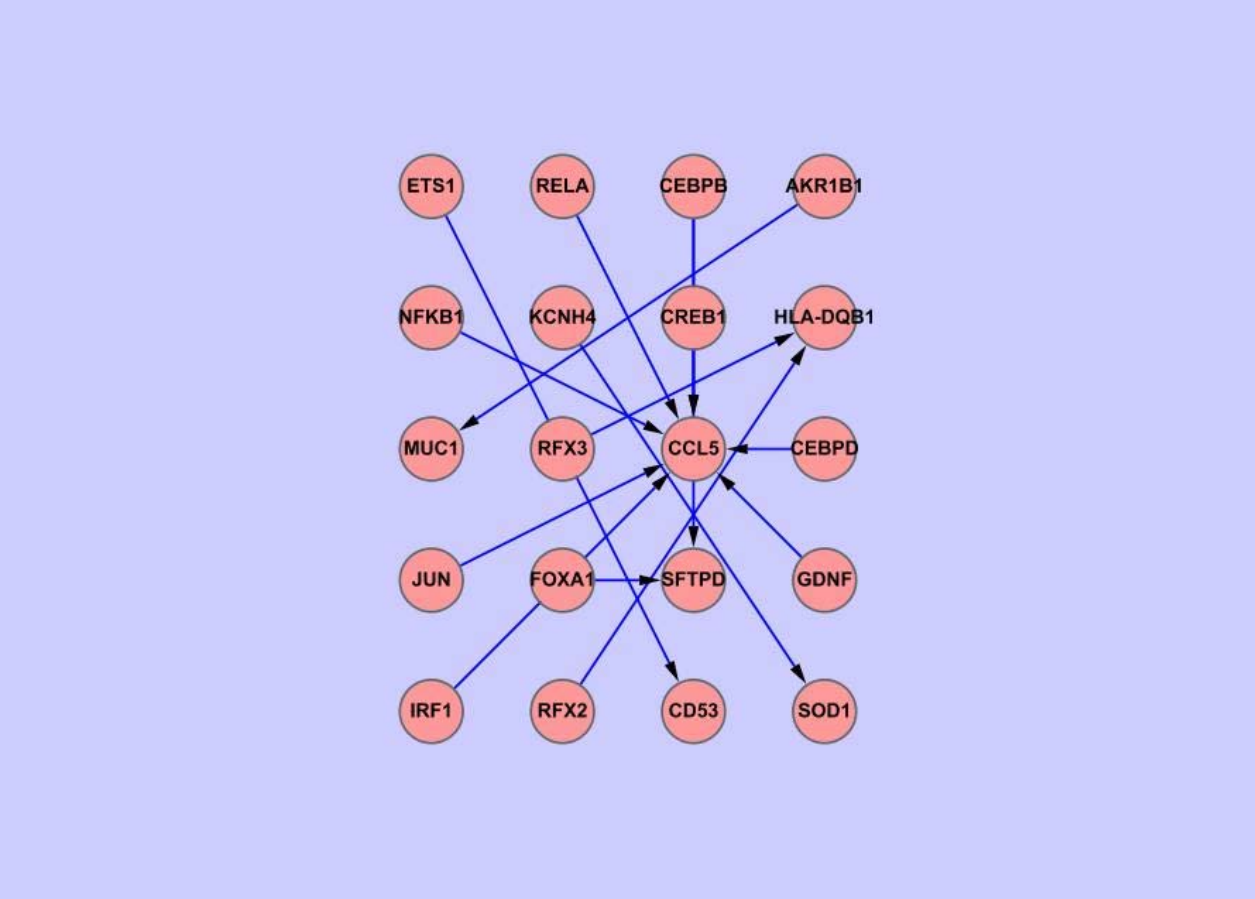


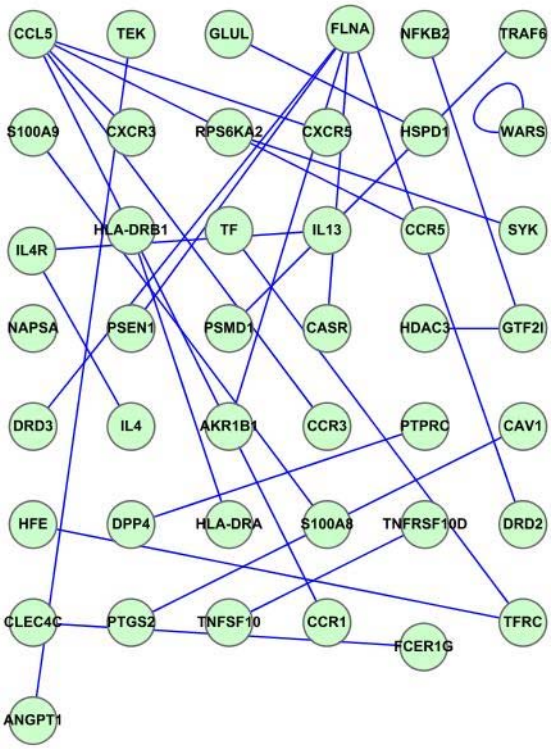
Type:Lung Cancer Subtype:Lung Adenocarcinoma



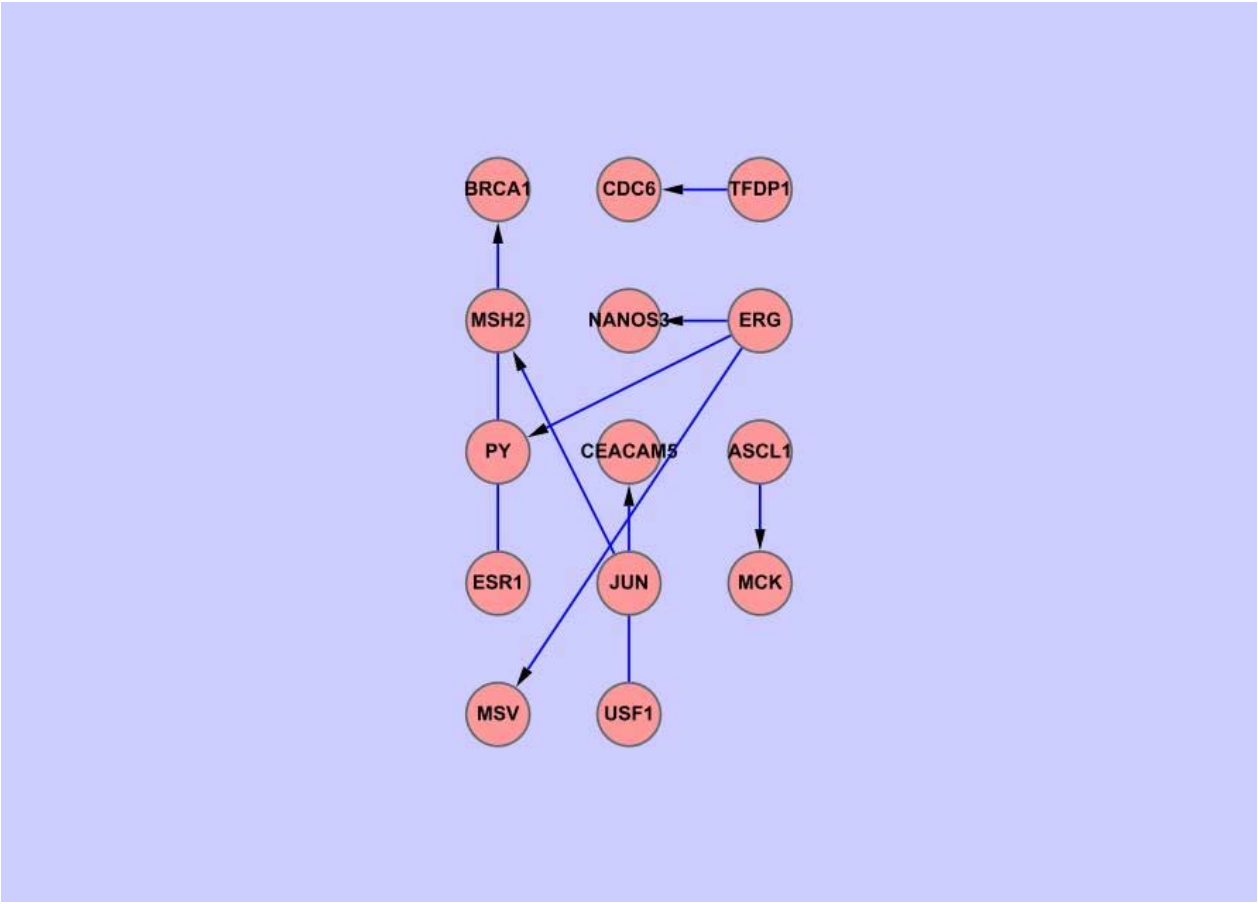


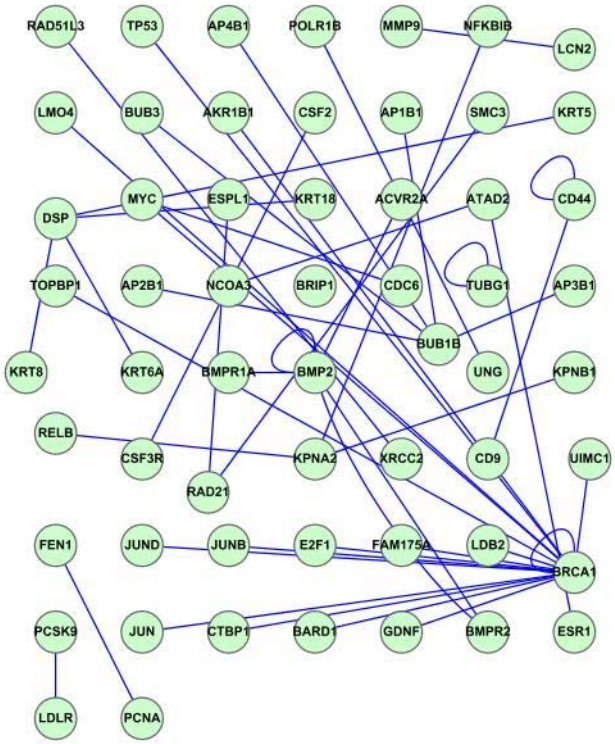
Type:Lung Cancer Subtype:Lung Carcinoid Tumor



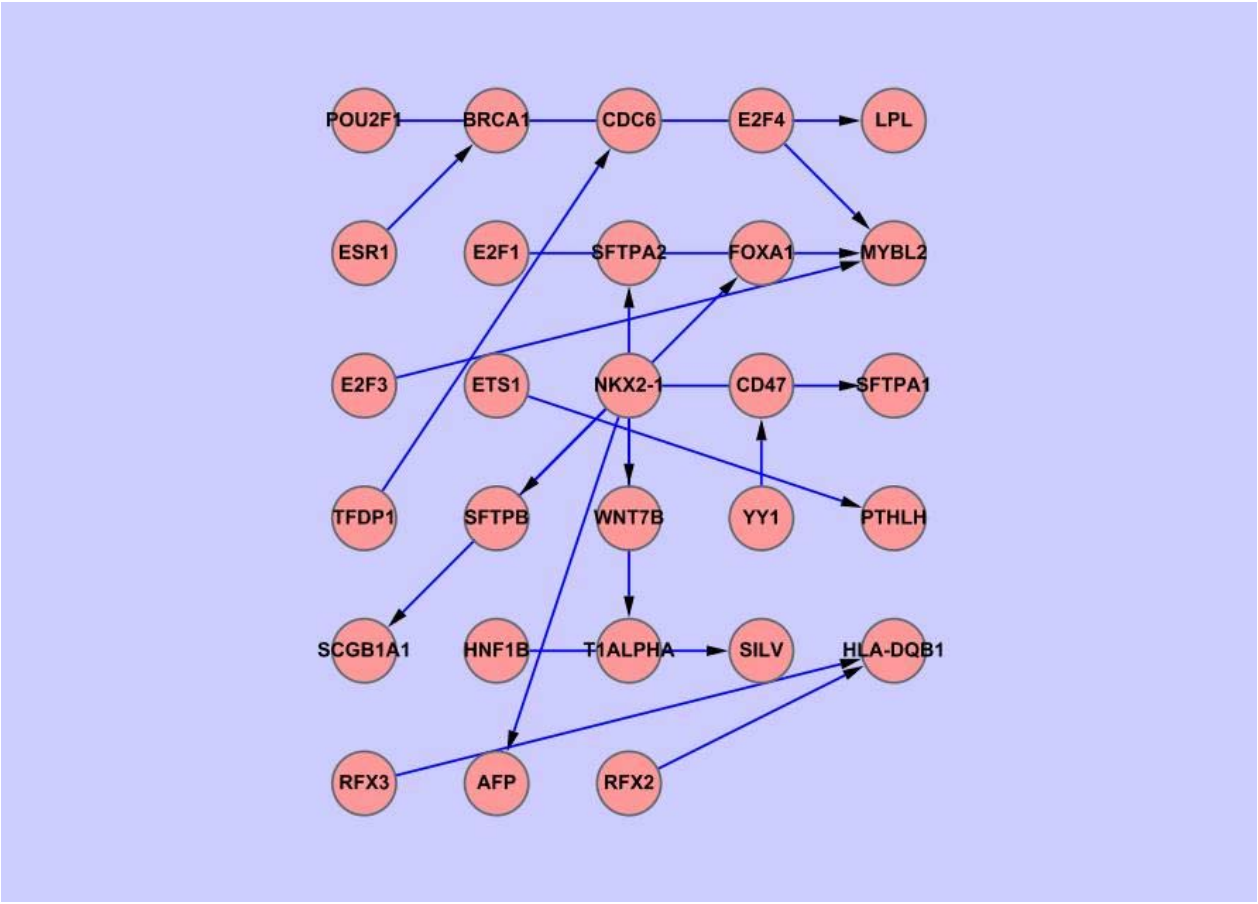


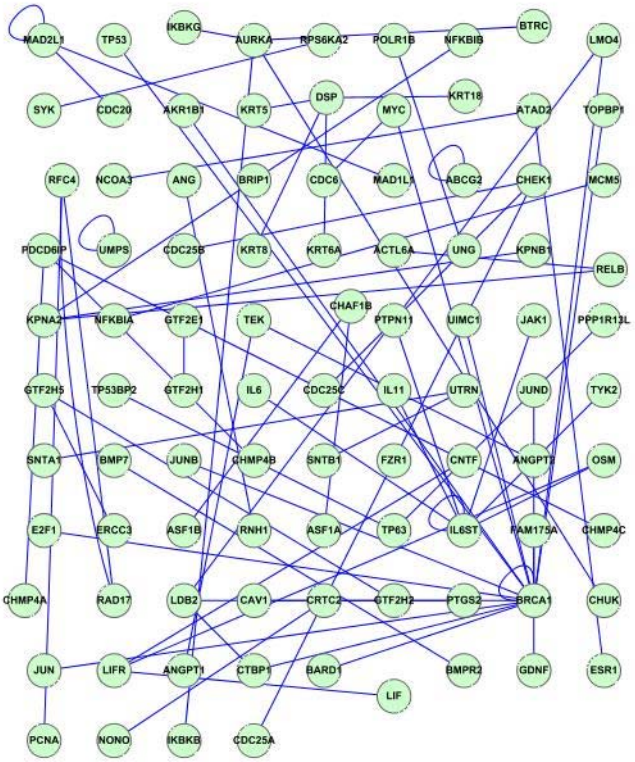
Type:Lung Cancer Subtype:Small Cell Lung Carcinoma





Type:Lung Cancer Subtype:Squamous Cell Lung Carcinoma





出 席 會 議 報 告 表

會議名稱	International Conference on Wavelet Analysis and Pattern Recognition	會議時間	2008 年 8 月 29 日 15 時 00 分
會議地點	香港		
會議主持人		出席人員	張培均
會議內容	這次的會議是以共同作者的身份前往，發表的論文是模糊理論在分類上的應用，在 Pattern Recognition 的領域裡，這是一個新的議題，我們提出利用模糊積分來整合分類特徵，以達到最佳分類精確度的演算法，這個方法目前廣泛應用在病毒分類、病毒宿主預測、生醫訊號的時間序列分類以及 DNA 微陣列的癌症分類上。在與會過程中，經過相關的討論，我們瞭解到關於模糊測度的估計方法，與被預測系統的特性具高度敏感性，分類特徵的選取在高度非線性系統往往不能以 Pearson's 係數作為模糊測度，到目前為止，這依舊是一個沒有定論的議題。以模糊積分來衡量不同估計特徵量之間的交互作用，是我未來研究癌症相關生化路徑網路交互作用的主要方向。 與會三天中，聽取眾多學者的論文發表，獲益良多。		

A NOVEL CLASSIFIER FOR INFLUENZA A VIRUSES BASED ON SVM AND LOGISTIC REGRESSION

Hsiang-Chuan Liu[1], Shin-Wu Liu[2], Pei-Chun Chang[1]
Wen-Chun Huang[3], Chien-Hsiung Liao[1]

[1]Department of Bioinformatics, Asia University, Taiwan
[2] National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA
[3]Graduate Institute of Educational Measurement and Statistics, Taichung University, Taiwan
E-MAIL: lhc@asia.edu.tw, liushin@mail.nih.gov, pcchang@asia.edu.tw,
huang.hwc@gmail.com, ace.liao@gmail.com

Abstract:

In search of good classifier of hosts of influenza A viruses is an important issue to prevent pandemic flu. The hemagglutinin protein in the virus genome is the major molecule that determining the range of hosts. In this paper, a novel classification algorithm of hemagglutinin proteins integrating SVM and logistic regression based on 4 kinds of Hurst exponents for each protein sequence is proposed. This method not used before is the first one integrating the physicochemical properties, fractal property, SVM and logistic regression classifier. For evaluating the performance of this new algorithm, a real data experiment by using 5-fold Cross-Validation accuracy is conducted. Experimental result shows that this new classification algorithm is useful and batter than SVM and logistic regression, respectively.

Keywords:

Influenza A viruses; Hurst exponent; SVM; logistic regression; SVM-Logistic regression

1. Introduction

Influenza A viruses are negative-strand RNA viruses that infect a wide variety of animals in the nature. The infection of human may cause significant mortality and morbidity worldwide [1]. The hemagglutinin (HA) protein in the virus genome is the major molecule that determining the range of hosts. The natural reservoir of influenza virus such as avian flu may emerges in strains infectious to human by mutation of HA protein and brings pandemic flu, therefore, in search of good classification algorithm of HA proteins is an important issue to prevent pandemic flu. In this paper, a novel classification algorithm of HA proteins combining Hurst exponents, SVM and logistic regression is proposed [2], [3], [4], [5]. This method not used before is the first one integrating the physicochemical properties, fractal property, support vector machine (SVM) and logistic regression classifier.

The protein residues were coded according to its physicochemical quantities of acidity, Van der waal volume, surface area and hydrophobicity in the situation of single amino acid [6], [7]

First step, the HA sequence data of serotype H5 of influenza A viruses with two classes used in this research were downloaded from public databases: Influenza Sequence Database (http://www.flu.lanl.gov). The sample included 90 HA protein sequences of human infections and 90 HA protein sequences of bird infections.

Second step, to replace each residue of amino acid in the sequences of the HA proteins with 4 physicochemical quantities.

Third step, computing the Hurst exponents of each non-symbolic sequences of the HA proteins, we can obtained four features of Hurst exponents in each sequences of the HA protein [2], [6], [7]..

Last step, two well known and appealing classifiers, Support Vector Machine (SVM) and Logistic regression (LR), and our new hybrid classifier combining SVM and LR were used to discriminate the correct class of the 180 HA proteins with four features of Hurst exponents

For evaluating the performance of above three classifiers, the above HA proteins data experiment by using 5-fold Cross-Validation accuracy is conducted.

This paper is organized as followings: four physicochemical quantities of 20 amino acids are introduced in section 2, Hurst exponent is introduced in section 3, support vector machine classifier is introduced in section 4, logistic regression is introduced in section 5, the new hybrid classifier combining SVM and logistic regression is introduced in section 6, experiment and result

are described in section 7 and final section is for conclusions and future works.

2. Four physicochemical properties of amino acids

There are four physicochemical quantities of acidity, Van der waal volume, surface area and hydrophobicity in the situation of single amino acid showed as Table 1 [2], [3]

Table 1.
20 amino acids and its 4 physicochemical quantities

Amino acid	Acidity	Van der waal Volume	Surface area	HydroPh-obicity
A	7.0	67	115	0.616
C	3.4	86	135	0.680
D	3.9	67	150	0.028
E	4.1	109	190	0.043
F	7.0	135	210	1.000
G	7.0	48	75	0.501
H	6.0	118	195	0.165
I	7.0	124	175	0.943
K	10.5	135	200	0.283
L	7.0	124	170	0.943
M	7.0	124	185	0.738
N	7.0	148	160	0.236
P	7.0	90	145	0.711
Q	7.0	114	180	0.251
R	12.5	167	225	0.000
S	7.0	73	115	0.359
T	7.0	93	140	0.450
V	7.0	105	155	0.825
W	7.0	163	255	0.878
Y	10.5	141	230	0.880

3. Hurst exponent

The Hurst exponent occurs in several areas of applied mathematics, including fractals and chaos theory, long term memory processes and spectral analysis [8]. Hurst exponent estimation has been applied in areas ranging from biophysics to computer networking. Estimation of the Hurst exponent was originally developed in hydrology. However, the modern techniques for estimating the Hurst exponent comes from fractal mathematics.

Estimating the Hurst exponent for a data set provides a measure of whether the data is a pure random walk or has underlying trends.

The Hurst exponent (H) is a statistical measure used to classify time series. H=0.5 indicates a random series while H>0.5 indicates a trend reinforcing series. The larger the H value is, the stronger the trend. Experiments with backpropagation Neural Networks show that series with large Hurst exponent can be predicted more accurately than those with H value close to 0.50. Thus the Hurst exponent provides a measure for predictability.

Three methods were used most often for the estimation of the Hurst exponent: the R/S method, the roughness–length (R–L) method and a variogram. The R/S method (Hurst et al., 1965) [9] is commonly perceived as the most suitable for the time series analysis, because it presents the relationship between irregular (singular) rescaled ranges, signal value and their local statistical properties relative to the scale factor.

In this study R/S method is used. R/S method [10] is based on empirical observations by Hurst and estimates H are based on the R/S statistic. It indicates (asymptotically) second-order self-similarity. H is roughly estimated through the slope of the linear line in a log-log plot, depicting the R/S statistics over the number of points of the aggregated series. That is, given a time sequence of observations, w_t define the Series

$$W(t,\tau)=\sum_{u=1}^{t}(w_u-\overline{w}_\tau),1\le t\le\tau, \quad (5)$$

Where $\overline{w}_\tau=\frac{1}{\tau}\sum_{t=1}^{\tau}w_t$. (6)

Define

$$R(\tau)=\max_{t=1}^{\tau}W(t,\tau)-\min_{t=1}^{\tau}W(t,\tau) \quad (7)$$

and $$S(\tau)=\sqrt{\left(\frac{1}{\tau}\sum_{t=1}^{\tau}(w_t-\overline{w}_\tau)^2\right)} \quad (8)$$

In plotting $\log\frac{R(\tau)}{S(\tau)}$ against $\log\tau$, we expect to get a line whose slope determines the Hurst exponent.

4. Support vector machine (SVM), [11~14],

Given the training set of instance-labeled pairs $(\underline{x}_i,y_i),i=1,2,...,N$, where

$$\underline{x}_i\in R^n,y_i\in\{1,-1\},i=1,2,...,N \quad (9)$$

The support vector machine (SVM) algorithm (Boser,

Guyon, and Vapnik 1992 [11], Cortes and Vapnik 1995 [12]) requires

$$\min_{\underline{w},b,\xi} \frac{1}{2}\underline{w}'\underline{w} + c\sum_{i=1}^{N}\xi_i$$
$$\text{subject to } y_i\left(\underline{w}'\phi\left(\underline{x}_i\right)+b\right) \geq 1-\xi_i,$$
$$\xi_i \geq 0, \tag{10}$$
$$\textit{where} \quad b,c \in R, \underline{w}, \phi\left(\underline{x}_i\right) \in R^m$$
$$\phi : R^n \to R^m$$

For any testing point $\underline{x}_i \in R^n, y_i \in \{1,-1\}$, we can make an assignment according to the following formula.

$$d(\underline{x}_i) = \left[\underline{w}'\varphi\left(\underline{x}_i\right)+b-\left(1-\xi_i\right)\right]$$
$$y_i = \begin{cases} +1, & \text{if } d(\underline{x}_i) \geq 0 \\ -1, & \text{if } d(\underline{x}_i) < 0 \end{cases} \tag{11}$$

5. Multiple Logistic regression classifier

5.1 Multiple logistic regression model [4], [5]

Let $\left(x_{i1}, x_{i2}, ..., x_{in} y_i\right), i = 1,2,...,N$ be a sample data, satisfying $\underline{x}_i = \left(x_{i1}, x_{i2}, ..., x_{in}\right) \in R^n$, $y_i \in \{0,1\}$,

$$Y_i^{\perp\perp} \sim B\left(1, p_i\right), i = 1,2,...,N \tag{12}$$

The multiple logistic regression model is denoted as follows

$$P_i = P\left(Y_i = 1 \mid x_i\right) = \frac{1}{1+\exp\left[-\left(\alpha+\underline{\beta}'\underline{x}\right)\right]}, i = 1,2,...,N \tag{13}$$

where $\underline{\beta}' = \left(\alpha, \beta_1, \beta_2, ..., \beta_n\right)$ are parameters vector of regression coefficients.

5.2 Multiple logistic regression classifier [5]

We can obtain the likelihood function and log likelihood function as following equations (14) and (15)

$$L\left(p_1, p_2, ..., p_N\right) = \prod_{i=1,2,...,N} p_i^{y_i}\left(1-p_i\right)^{1-y_i} \tag{14}$$

$$l = \log L\left(p_1, p_2, ..., p_N\right) = \sum_{i=1}^{N}\left[y_i \log p_i + \left(1-y_i\right)\left(1-\log p_i\right)\right] \tag{15}$$

And we can get

$$l = l\left(\alpha, \underline{\beta}\right) = \sum_{i=1}^{N}\left[y_i \log p_i + \left(1-y_i\right)\left(1-\log p_i\right)\right]$$
$$= -\sum_{i=1}^{N}\left[\log\left(1+exp\left[-\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]\right)+\left(1-y_i\right)\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right] \tag{16}$$

Where $\alpha \in R, \underline{\beta}' = \left(\beta_1, \beta_2, ..., \beta_n\right) \in R^n$

Using Newton-Raphson's iterative algorithm, we can get the estimated regression coefficients of the multiple logistic regression model and the estimated multiple logistic regression equation as follows:

$$\hat{P}_i = \hat{P}\left(Y_i = 1 \mid x_i\right) = \frac{1}{1+\exp\left[-\left(\hat{\alpha}+\underline{\hat{\beta}}'\underline{x}\right)\right]} \tag{17}$$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}_{k+1} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}_k - \begin{bmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta_1} & \cdots \frac{\partial^2 l}{\partial \alpha \partial \beta_n} \\ \frac{\partial^2 l}{\partial \beta_1 \partial \alpha} & \frac{\partial^2 l}{\partial \beta_1^2} & \cdots \frac{\partial^2 l}{\partial \alpha \partial \beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \beta_n \partial \alpha} & \frac{\partial^2 l}{\partial \beta_n \partial \beta_1} & \cdots \frac{\partial^2 l}{\partial \beta_n^2} \end{bmatrix}_k^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \beta_1} \\ \vdots \\ \frac{\partial l}{\partial \beta_n} \end{bmatrix}_k \tag{18}$$

where

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{N}\left[\frac{1}{1+\exp\left[-\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]} - \left(1-y_i\right)\right] \tag{19}$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N}\left[\frac{1}{1+\exp\left[-\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]} - \left(1-y_i\right)\right] x_{ij}, \; j = 1,2,...,n \tag{20}$$

$$\frac{\partial^2 l}{\partial \alpha^2} = -\sum_{i=1}^{N}\frac{\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)}{\left[1+\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]^2} \tag{21}$$

$$\frac{\partial^2 l}{\partial \beta_j^2} = -\sum_{i=1}^{N}\frac{x_{ij}^2 \exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)}{\left[1+\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]^2}, \; j = 1,2,...,n \tag{22}$$

$$\frac{\partial^2 l}{\partial \alpha \partial \beta_j} = \frac{\partial^2 l}{\partial \beta_j \partial \alpha} = -\sum_{i=1}^{N}\frac{x_{ij} \exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)}{\left[1+\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]^2}, \; j = 1,2,...,n \tag{23}$$

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}=\frac{\partial^2 l}{\partial\beta_k\partial\beta_j}=-\sum_{i=1}^{N}\frac{x_{ij}x_{ik}\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)}{\left[1+\exp\left(\alpha+\underline{\beta}'\underline{x}_i\right)\right]^2},\ j,k=1,2,...,n \qquad P_i=P\left(Y_i=1\mid d\left(\underline{x}_i\right)\right)=\frac{1}{1+\exp\left[-\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]},\ i=1,2,...,N \tag{29}$$

Increment k; until $\left\|\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_n\end{bmatrix}_{k+1}-\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_n\end{bmatrix}_{k}\right\|<\varepsilon$ (24)

Then $y_i=\begin{cases}1 & if\ \hat{P}_i=\hat{P}\left(Y_i=1\mid x_i\right)\ge 0.5\\ 0 & if\ \hat{P}_i=\hat{P}\left(Y_i=1\mid x_i\right)<0.5\end{cases}$ (25)

6. SVM-Logistic regression classifier

In this paper, an improved hybrid classifier combining SVM and logistic regression is proposed here.
First step, using the SVM classifier, we can find the signed distance, $d(\underline{x}_i)$, between the point $\underline{x}_i=\left(x_{i1},x_{i2},...,x_{in}\right)$ and the hyperplane in SUM,
Second step, to consider the sample data $\left(d\left(\underline{x}_i\right),y_i\right),i=1,2,...,N$, using the simple logistic regression to classify y_i .

6.1 Mathematical formulas

Let $\left(x_{i1},x_{i2}....,x_{in}y_i\right),i=1,2,...,N$ be a sample data, satisfying $\underline{x}_i=\left(x_{i1},x_{i2},...,x_{in}\right)\in R^n,\ y_i\in\{0,1\}$, (26)
Using the above support vector machine (SVM) algorithm, from equation (11), for any point $\underline{x}_i\in R^n$, we can obtain the signed distance as below

$$d(\underline{x}_i)=\left[\underline{w}'\varphi\left(\underline{x}_i\right)+b-\left(1-\xi_i\right)\right] \tag{27}$$

6.2 Simple logistic regression classifier of the working sample data

Let the working sample data $\left(d\left(\underline{x}_i\right),y_i\right),i=1,2,...,N$ satisfying $d\left(\underline{x}_i\right)\in R,\ y_i\in\{1,0\}$

$$Y_i^{\perp\perp}\sim B\left(1,p_i\right),i=1,2,...,N \tag{28}$$

The simple logistic regression model is denoted as follows

Similarly as multiple logistic regression classifier, we can get log likelihood function, the estimated regression coefficients of the simple logistic regression model and the estimated simple logistic regression equation as follows:

$$l=l(\alpha,\beta)=\sum_{i=1}^{N}\left[y_i\log p_i+\left(1-y_i\right)\left(1-\log p_i\right)\right]$$
$$=-\sum_{i=1}^{N}\left[\log\left(1+exp\left[-\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]\right)+\left(1-y_i\right)\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right] \tag{30}$$

$$\hat{P}_i=\hat{P}\left(Y_i=1\mid d\left(\underline{x}_i\right)\right)=\frac{1}{1+\exp\left[-\left(\hat{\alpha}+\hat{\beta}d\left(\underline{x}_i\right)\right)\right]} \tag{31}$$

$$\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}_{k+1}=\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}_{k}-\begin{bmatrix}\frac{\partial^2 l}{\partial\alpha^2} & \frac{\partial^2 l}{\partial\alpha\partial\beta}\\ \frac{\partial^2 l}{\partial\beta\partial\alpha} & \frac{\partial^2 l}{\partial\beta^2}\end{bmatrix}_k^{-1}\begin{bmatrix}\frac{\partial l}{\partial\alpha}\\ \frac{\partial l}{\partial\beta}\end{bmatrix}_k \tag{32}$$

where

$$\frac{\partial l}{\partial\alpha}=\sum_{i=1}^{N}\left[\frac{1}{1+\exp\left[-\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]}-\left(1-y_i\right)\right] \tag{33}$$

$$\frac{\partial l}{\partial\beta}=\sum_{i=1}^{N}\left[\frac{1}{1+\exp\left[-\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]}-\left(1-y_i\right)\right]d\left(\underline{x}_i\right) \tag{34}$$

$$\frac{\partial^2 l}{\partial\alpha^2}=-\sum_{i=1}^{N}\frac{\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)}{\left[1+\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]^2} \tag{35}$$

$$\frac{\partial^2 l}{\partial\beta^2}=-\sum_{i=1}^{N}\frac{\left[d\left(\underline{x}_i\right)\right]^2\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)}{\left[1+\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]^2} \tag{36}$$

$$\frac{\partial^2 l}{\partial\alpha\partial\beta}=\frac{\partial^2 l}{\partial\beta\partial\alpha}=-\sum_{i=1}^{N}\frac{d\left(\underline{x}_i\right)\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)}{\left[1+\exp\left(\alpha+\beta d\left(\underline{x}_i\right)\right)\right]^2} \tag{37}$$

Increment k; until $\left\|\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}_1\end{bmatrix}_{k+1}-\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix}_{k}\right\|<\varepsilon$ (38)

Then $y_i=\begin{cases}1 & if\ \hat{P}_i=\hat{P}\left(Y_i=1\mid d\left(\underline{x}_i\right)\right)\ge 0.5\\ 0 & if\ \hat{P}_i=\hat{P}\left(Y_i=1\mid d\left(\underline{x}_i\right)\right)<0.5\end{cases}$ (39)

7. Experiment and result

The sequence data of serotype H5 of Influenza A

viruses with two classes used in this research were obtained from public databases: Influenza Sequence Database (http://www.flu.lanl.gov). The sample included 90 HA protein sequences of human infections and 90 HA protein sequences of bird infections.

The protein residues were coded according to its physicochemical quantities of acidity, Van der waal volume, surface area and hydrophobicity in the situation of single amino acid as Table 1.

Computing the Hurst exponents of each non-symbolic sequences of the HA proteins, we can obtain four features represented as Hurst exponents respectively in each sequences of the HA protein.

The above real data with four features in terms of Hurst exponents is applied to evaluate the performances of the Support Vector Machine (SVM) algorithm, logistic regression and the proposed classifier combining SVM and logistic regression classifier by using 5-fold Cross-Validation method to compute the accuracies of the response category variable.

The experimental results for Accuracies of above three classifiers are listed in Table 2. We can find that our new classification algorithm is useful and batter than SVM and logistic regression, respectively.

Table 2 Accuracies of three classifiers

Classifier	5-fold CV accuracy
SVM	0.8056
LR	0.8833
SVM-LR	0.9000

8. Conclusions and future works

In search of good classifier of influenza viruses is an important issue to prevent pandemic flu. In this paper, a novel classification algorithm of HA proteins integrating SVM and logistic regression based on 4 kinds of Hurst exponents for each protein sequence is proposed. This method not used before is the first one integrating the physicochemical properties, fractal property, SVM and logistic regression classifier. For evaluating the performance of this new algorithm, a real data experiment by using 5-fold Cross-Validation accuracy is conducted. Experimental result shows that this new classification algorithm is useful and batter than SVM and logistic regression, respectively.

Our proposed new classifier can be used to classify not only the data of Influenza A viruses but also the data of other biological sequences.

In future, we will consider look for some further improving classification algorithms by using Hurst exponent and other hybrid Classifiers.

Acknowledgements

This paper is partially supported by the National Science Council grant (NSC 96-2413--H-468-001).

References

[1] P. Pale, " Influenza: old and new threats", Nat. Med , Vol.10, pp. 82–87, 2004.

[2] H. E. Hurst, "Long term storage capacity of reservoirs", Transactions of the American Society of Civil Engineers 116, pp. 770-799, 1951.

[3] C. Cortes, and V., Vapnik, "Support-vector network" , Machine Learning, Vol. 20, pp. 273-297, 1995.

[4] D. R. Cox, and E. J. Snell, The analysis of binary data (2nd ed.) London, Chapman & Hall, 1989.

[5] Hsiang-Chuan Liu, Yu-Du Jheng, Guey-Shya Chen, Bai-Cheng Jeng, "A new classification algorithm combining Choquet integral and logistic regression", 2008 International Conference on Machine Learning and Cybernetics, 12-15 July 2008 Kunming, China (accepted).

[6] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers., and Y. Kawaoka, "Evolution and ecology of influenza A viruses", Microbiol. Rev., Vol. 56, pp. 152-179, 1992.

[7] N. J. Cox and K. Subbarao, "Global epidemiology of influenza: Past and present", Annu. Rev. Med., Vol. 5*1*, pp. 407-421, 2000.

[8] T. Di Matteo, T. Aste and M. M. Dacorogna, "Long term memories of developed and emerging markets: using the scaling analysis to characterize their stage of development", Journal of Banking & Finance 29/4, pp. 827-851, 2005.

[9] H. E. Hurst, R. Black, Y. M. Sinaika, "Long term storage capacity of reservoirs" , An experimental study Constable, London, 1965.

[10] Roger Kalden, Sami Ibrahim, "Searching for Self-Similarity in GPRS", PAM, pp. 83-92, 2004.

[11] B.E. Boser, I.M. Guyon, and V. Vapnik, "A training

algorithm for optimal margin classifiers", In Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, 1992. ACM.

[12] C. Cortes, and V. Vapnik, "Support-vector network", Machine Learning, Vol. 20, pp. 273-297, 1995.

[13] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY. Springer-Verlay, 1995.

[14] C.-C. Chang, and C.-C. Lin, LIBSVM; a library for support vector machine Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2004

出 席 會 議 報 告 表

會議名稱	International Conference of Nature Computatin	會議時間	2008 年 10 月 19 日 10 時 20 分
會議地點	中國山東省濟南市南郊賓館		
會議主持人	Zu-Guo Yu	出席人員	Zu-guo Yu, Chung-Hung Li, Mao-Zu Guo, Ben-Zheng Wei, Jun Wang Yan Xu 等
會議內容	學生參加的 session 為 Bioinformatics and Bio-medical Engineering。本 session 共 7 篇論文。 1.Phylogenetic analysis of Polyomaviruses based on their complete genomes 是由 Zu-guo Yu 所報告。本篇論文是分析 70 個 genomes corresponding to nine mammalian，與 two avian。 2.Physiochemical Constrains in Influenza A Hemagglutinin 是由學生所報告。本論文是用 Fuzzy measure 與 Choquet 去檢測 HA protein 3 種不同物理化學性質的感染相關性，最後由 Hurst exponent 的 H value 做分類依據 3.prediction of protein-protein Interactions from Secondary Structure in Binding Motifs Using the Statistic Method 是由 Mao-Zu Guo 所報告。本論文是用幾個統計方法(cross-validation, false-positive) 去看在 protein-protein interaction binding motif region 上 helice, sheets,與 disordered secondarystructure 的 frequence 當門檻值然後再去預測 protein-protein interaction。 4.Protein structure classification using local Holder exponents estimated by wavelet transform 是由 Zu-Guo Yu 所報告。本論文是使用 local Holder exponent 去辨識蛋白質序列。 5.SSART Being Applied to Approaching Accurate Value of High Frequency Field 是由 Ben-Zheng Wei 所報告。 6.TagSNP Selection Using Maximum Density Subgraph 是由 Jun Wang 所報告。本報告是利用新的方法(combine clustering 與 graph algorithm)找出 small subset of informative SNP(tagSNP)。 7.The Research on the Relationship between Opioid Receptors 是由 Yan Xu 所報告。		

Physiochemical Constraints in Influenza A Hemagglutinin

Jiunn-I Shieh
Department of Information Science and Applications
Asia University, Taichun, Taiwan, R.O.C.
jishieh@asia.edu.tw

Kuei-Jen Lee
Department of Health and Nutrition Biotechnology
Asia University, Taichun, Taiwan, R.O.C.
kjlee@asia.edu.tw

Jing-Doo Wang
Department of Computer Science and Information Engineering
Asia University, Taichun, Taiwan, R.O.C.
jdwang@asia.edu.tw

I-Chun Chen, Am-Chou Chen, Pei-Chun Chang, and Hsiang-Chuan Liu
Department of Bioinformatics
Asia University, Taichun, Taiwan, R.O.C.
mmisssoul@gmail.com, gg.village@gmail.com, {pcchang,lhc}@asia.edu.tw

Abstract

Influenza A viruses are negative-strand RNA viruses. The gene of hemagglutinin (HA) protein in the virus genome is the major molecule that determining the range of hosts. Mutation of HA gene may bring infection cross species. In this paper, we studied physicochemical constraints during the variations of HA gene. Fuzzy measure and Choquet integral were used to estimate the combining effect of different physicochemical properties for single residue in HA protein that related to infective events. With this method, a HA sequence was quantified residue by residue and produced a value series. Finally, the Hurst exponent was adopted to infer the constraints in the series. We found that the physicochemical constraints in HA sequences mainly falling into two classes of interdependence strength during gene variation, that were distinct from the diversity of clusters in the phylogenetic analysis.

4. Introduction

Influenza A viruses are negative-strand RNA viruses that infect a wide variety of animals in the nature. The infection of human may cause significant mortality and morbidity worldwide [1]. The gene of hemagglutinin (HA) protein in the virus genome is the major molecule that determining the range of hosts. The natural reservoir of influenza virus such as avian flu may emerge in strains infectious to human by mutation of HA gene [2,3]. Owing to that, it is important to understand the variation nature of HA gene. In the past, the researches in this field mainly have been focused on the phylogenetic reconstructions [4,5]. As shown in the explosive information on HA sequences, the reconstruction of a phylogenetic tree can provide abundant evolution information, and help in understanding the drifts of influenza hosts [6]. However, the feature and tendency about physicochemical properties of gene variations for specific host are never been discussed.

Fuzzy measure theory considers a number of special classes of measurements, each of which is characterized by a special property. In the fuzzy measure theory, the conditions are precise, but the information about an element alone is insufficient to determine which special classes of measure should be used. The fuzzy measure estimates the possible interactions among the special classes of measurements [7]. Choquet integral is a tightly related concept with fuzzy measure. It assesses the integrated effect for some issue based on the concept of fuzzy measure [7,8]. The Hurst exponent (H) is a statistical measure used to classify time series [9]. For example, H=0.5 indicates a random series while H>0.5 indicates a constrained reinforcing series. The larger the H value is, the stronger the constraint. In this paper, we studied the physicochemical constraints of HA protein of Influenza A

viruses regarding to serotypes H1, H3, and H5. We concerned three types of physicochemical property for each residue that have acidity, Van der waal volume, and hydrophobicity [10]. Pearson's correlation coefficient was used to quantify the dependence of physicochemical properties on infection hosts, human or avian. For each residue, there were three values of Pearson's correlation coefficient corresponding to three types of physicochemical properties. Based on the coefficients, Sugeno λ-measure [11] was adopted to calculate the fuzzy measure. Subsequently, the Choquet integral was applied to assess the integrated effect of physicochemical properties on infection hosts for each residue. A protein sequence implies a series of integral values. Finally, we used Hurst exponent to analyze the value series for exploring the integrated physicochemical constraints in the protein sequence.

5. Methods

5.1. Sequence data collection

The sequence data of Influenza A viruses used in this research were obtained from public databases: Influenza Sequence Database (http://www.flu.lanl.gov). All HA nucleotide sequences of human and birds in this databases were downloaded on October 16, 2006. The HA sequences were extracted, of which less than 900 nucleotides were considered as partial sequences and were excluded from this study. Identically coded sequences are considered as duplicates and only the earliest isolated strain among the duplicates was used as a representative sequence in the group. In total, we had 831 H1 sequences, 3018 H3 sequences and 1376 H5 sequences for our analysis. All sequences were isolated between 1963 and 2006 from locations around the globe. The exact isolation time (calendar year), host type and location can be found in the strain names.

2.2. Residue coding

The sequence alignment processes were implemented in ClustalX 3.14 [12] regarding to H1, H3, and H5. After alignment, the sequence length regarding to H1, H3, and H5 were 565, 567, and 583 amino acids respectively. The protein residues were coded according to its values of acidity, Van der waal volume, and hydrophobicity in the situation of single amino acid [10, 13], as shown in table 1. For every physicochemical property, we had a matrix size of 831x565 for H1 group, 3018x567 for H3 group, and 1376x583 for H5 group.

Table 1. The residue codes regarding to acidity, Van der waal volume, and hydrophobicity.[a]

Amino acid	Acidity	Van der waal volume	Hydrophobicity
Alanine	7.0	67.	0.616
Cysteine	8.4	86.	0.68
Aspartic acid	3.9	67.	0.028
Glutamic acid	4.1	109.	0.043
Phenylalanine	7.0	135.	1.
Glycine	7.0	48.	0.501
Histidine	6.0	118.	0.165
Isoleucine	7.0	124.	0.943
Lysine	10.5	135.	0.283
Leucine	7.0	124.	0.943
Methionine	7.0	124.	0.738
Asparagine	7.0	148.	0.236
Proline	7.0	90.	0.711
Glutamine	7.0	114.	0.251
Arginine	12.5	167.	0.
Serine	7.0	73.	0.359
Threonine	7.0	93.	0.45
Valine	7.0	105.	0.825
Tryptophan	10.5	163.	0.878
Tyrosine	7.0	141.	0.88

[a]The gaps in the aligned sequences were coded as 7., 0., and 0.5 for acidity, Van der waal volume, and hydrophobicity.

2.3. Inference of physicochemical constraints

Choquet integral is defined to integrate functions with respect to the fuzzy measure [7]. It is very useful in assessment of the effect that results from the nonlinear interactions. The definitions of fuzzy measure and Choquet integral are as follows:

Definition 1. Let N be a finite set of criteria. A discrete fuzzy measure on N is a set function $v: 2^N \to [0,1]$ which satisfies the following axioms:

(i) $v(\phi) = 0$, $v(N) = 1$ (boundary conditions)

(ii) $A \subseteq B$ implies $v(A) \le v(B)$ (monotonicity) for $A, B \in 2^N$.

For each subset of criteria $S \subseteq N$, $v(S)$ can be interpreted as the weight of the coalition S.

The Sugeno λ-measure is a special case of fuzzy measures. It has the following definition.

Definition 2. Let $N = \{X_1, X_2, \cdots, X_n\}$ be a finite set and $\lambda \in (-1, \infty)$. A Sugeno λ-measure is a function v from 2^N to [0, 1] with properties:

(i) $v(N) = 1$.

(ii) if $A, B \subseteq 2^N$ with $A \cap B = \phi$ then $v(A \cup B) = v(A) + v(B) + \lambda v(A) v(B)$.

As a convention, the value of v at a singleton set $\{X_i\}$ is called a density and is denoted by $v\{X_i\}$. In addition, we have that λ satisfies the property

$$\lambda + 1 = \prod_{i=1}^{n} (1 + \lambda v(X_i)) \quad (1)$$

Tahani and Keller [14] as well as Wang and Klir [15] have showed that that once the densities are known, it is possible to use the previous polynomial to obtain the values of λ uniquely.

Definition 3. Let v be a fuzzy measure on *N*. The discrete Choquet integral of function *x*: $N \rightarrow R$ with respect to v is defined by $C_v(x) = \sum_{i=1}^{n} x_{(i)} \left[v(A_{(i)}) - v(A_{(i+1)}) \right]$, where $(\cdot)$ indicates a permutation on *N* such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Also $A_{(i)} = \{x_{(i)}, \cdots, x_{(n)}\}$, and $A_{(n+1)} = \phi$. For instance, if $x_1 \leq x_3 \leq x_2$, then rank x_1, x_2, x_3 from the smallest one to the largest one. The result is $x_{(1)} = x_1$, $x_{(2)} = x_3$, $x_{(3)} = x_2$. Finally,

$$C_v(x_1, x_2, x_3) = x_1 * [v(\{x_1, x_2, x_3\})] + (x_3 - x_1) * [v(\{x_2, x_3\})] + (x_2 - x_3) * [v(\{x_2\})] \quad (2)$$

The discrete Choquet integral takes into account the interaction by means of the fuzzy measure v. If the criteria are independent, the fuzzy measure is additive. Then, the discrete Choquet integral coincides with the weighted arithmetic mean method. That is, $C_v(x) = \sum_{i=1}^{n} v(i) x_i$, $x \in R^n$. In this study, the correlation-based method proposed by Hsiang-Chuan Liu in 2006 [16,17] to construct the fuzzy measures in the discrete Choquet integral was used.

The Hurst exponent occurs in several areas of applied mathematics, including fractals and chaos theories, long memory processes and spectral analysis. Hurst exponent estimation has been applied in areas ranging from biophysics to computer networking. Estimation of the Hurst exponent was originally developed in hydrology. However, the modern techniques for estimating the Hurst exponent come from fractal mathematics.

Estimating the Hurst exponent for a data set provides a measure of whether the data is a pure random walk or has underlying trends. Another way to state this is that a random process with an underlying trend has some degree of autocorrelation. Furthermore, when the autocorrelation has a very long (or mathematically infinite) decay this kind of Gaussian process is sometimes referred to *as a long memory process.*

The Hurst exponent (H) is a statistical measure used to classify time series. H=0.5 indicates a random series while H>0.5 indicates a trend reinforcing series. The larger the H value is, the stronger the trend. In this paper we investigate the use of the Hurst exponent to classify series of financial data representing different periods of time. Experiments with back propagation Neural Networks show that series with large Hurst exponent can be predicted more accurately than those with H value close to 0.50. Thus the Hurst exponent provides a measure for predictability.

Three methods were used most often for the estimation of the Hurst exponent: the R/S method, the roughness–length (R–L) method and variogram. The R/S method [18] is commonly perceived as the most suitable for the time series analysis on the stock market or an optimal volume of water reservoirs, because it presents the relationship between irregular (singular) rescaled ranges, signal value and their local statistical properties relative to the scale factor. In this study R/S method is used. R/S method [19] is based on empirical observations by Hurst and estimates H are based on the R/S statistic. It indicates (asymptotically) second-order self-similarity. H is roughly estimated through the slope of the linear line in a log-log plot, depicting the R/S statistics over the number of points of the aggregated series. That is, given a time sequence of observations w_t, define the series $W(t,\tau) = \sum_{u=1}^{t} (w_t - \overline{w}_\tau)$, where $\overline{w}_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} w_t$.

Define $R(\tau) = \max_{t=1}^{\tau} W(t,\tau) - \min_{t=1}^{\tau} W(t,\tau)$

and $S(\tau)=\sqrt{\left(\frac{1}{\tau}\sum_{t=1}^{\tau}(w_t-\overline{w}_\tau)^2\right)}$. In plotting $\log\frac{R(\tau)}{S(\tau)}$ against $\log\tau$, we expect to get a line whose slope determines the Hurst exponent.

There is a 7-step to make hurst exponent analyze:

Step 1. With quantizing three properties each amino acid of each protein sequence, we have three time series for each protein sequence.

Step 2. For each property, normalize the data for each position which the same position of aligned protein sequences for affecting human and birds. That is, label elements in the sample by *l* and treat each position in aligned protein sequence as a random variable. *Assume the size of the sample is k. For the element l,* let *i*-th position of aligned protein sequences for property *m* be a random variable $X_i^{l,m}$ where $1\leqq l\leqq$ k, $1\leqq m\leqq 3$, and *n* is the length of aligned protein sequences. If $\max_l\{X_i^{l,m}\}-\min_l\{X_i^{l,m}\}$ ≠0, then $Z_i^{l,m}=\frac{X_i^{l,m}-\min_l\{X_i^{l,m}\}}{\max_l\{X_i^{l,m}\}-\min_l\{X_i^{l,m}\}}$. Otherwise, set $Z_i^{l,m}=0$.

Step 3. Let Y^l be a random variable which is 1 if affecting the human and 0 otherwise for the element *l*. Let $X_i^m=(X_i^{1,m},X_i^{2,m},\cdots,X_i^{k,m})'$ and $Y=(Y^1,Y^2,\cdots,Y^k)'$. For each *m*, compute corr(X_i^m, Y) where "corr" is the Pearson correlation coefficient. For each *m*, define the weight w_i^m to be $\frac{1+corr(X_i^m,Y)}{2}$ for each *i*. That is, $v(\{X_i^m\})=w_i^m$ for $1\leqq m\leqq 3$ *and* $1\leqq i\leqq n$.

Step 4. For using Sugeno λ-measure, solve (1) for λ. Then, for each *i* compute $v(\{X_i^1,X_i^2\})$, $v(\{X_i^1,X_i^3\})$, $v(\{X_i^2,X_i^3\})$ by Sugeno λ-measure. Note that $v(\{X_i^1,X_i^2,X_i^3\})=1$.

Step 5. Combined the three properties to be one, compute the Choquet integral for each position by equation (2). Then we get one time series for each aligned protein sequence.

Step 6. Calculate Hurst exponent for each aligned protein sequence.

Step 7. Analyze the results.

The above steps were calculated using Matlab package, except for Hurst exponent was obtained from the website: http://www.mathworks.com/matlabcentral/.

2.4. Results

We calculated the Hurst exponent regarding to H1, H3, and H5 to infer the physicochemical interdependency among the residues in the HA protein. The serotype H1 are shown in Fig.1, there are 2 clusters in the frequency distributions of Hurst exponents for human strains and avian strains. The Hurst exponent is nearby 1 for one cluster, and nearby 0.5 for another cluster. That mean some variations are constrained strongly, and some variations are random-like. The tendency of H3 is shown in Fig.2 and similar to H1, but the Hurst exponents in the two clusters are closer and away from 1 and 0.5. The results about H5 are shown in Fig.3, the distribution pattern is different from H1 and H3 for avian strains. There are three clusters in the frequency distribution.

The phylogenetic analysis is based on the mutation frequency between residues regarding homologous proteins. The evolution of quantitative property during the process of residue changes is ambiguous. In this study, we proposed a method based on the quantitative properties of residues regarding to infection issue of Influenza A viruses to estimate the constrain strength in the HA proteins. The distribution of constrain strength are distinct from the diversity of clusters in the phylogenetic analysis.

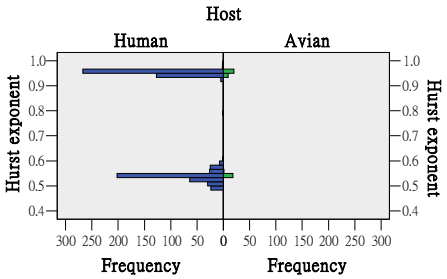


Figure 1. The frequency distribution of H1 Hurst exponents for human strains and avian strains.

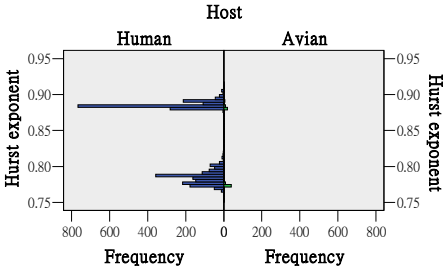


Figure 2. The frequency distribution of H3 Hurst exponents for human strains and avian strains.

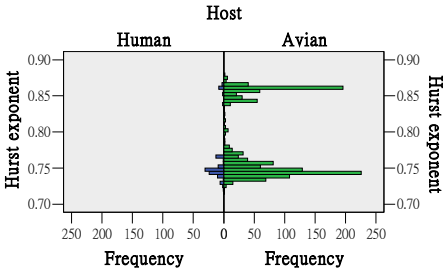


Figure 3. The frequency distribution of H5 Hurst exponents for human strains and avian strains.

2.5. Discussion

The gene of HA protein in the virus genome is the major molecule that determining the range of hosts. Basically, the infection process is physicochemical interaction between receptor of host and HA protein. For the sake of successful infection, the gene variations must follow certain rules under physicochemical base. Higher value of Hurst exponent implies more constraints or intra-structure in the sequence properties. As to that, the gene variations are apt to destroy the intra-structure with high value of Hurst exponent. The variation tolerance is different for the same serotype of HA corresponding to the different clusters of Hurst exponents.

4. Conclusions

The constraints in HA sequences mainly fall into two classes of Hurst strength during gene variations. That imply the variation tolerance of HA gene is diverse in the same serotype of HA.

Acknowledgements

This work was supported by the National Science Council, grant no. NSC 95-2221-E-468-006-.

References

[1] P. Palese. Influenza: old and new threats. *Nat. Med., Vol 10, pp. s82–s87, 2004.*

[2] R.G. Webster, W.J. Bean, O.T. Gorman, and T.M. Chambers, Y. Kawaoka., Evolution and ecology of influenza A viruses. *Microbiol. Rev., Vol 56, pp. 152–179, 1992.*

[3] N.J. Cox and K. Subbarao. Global epidemiology of influenza: Past and present. *Annu. Rev. Med., Vol 51, pp. 407–421, 2000.*

[4] W.M. Fitch, R.M. Bush, C.A. Bender, and N.J. Cox. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Nat. Acad. Sci., Vol 94, pp. 7712–7718, 1997.*

[5] R.M. Bush, W.M. Fitch, C.A. Bender, and N.J. Cox. Positive selection on the H3 hemagglutiningene of human influenza virus A. *Mol. Biol. Evol., Vol 16, pp. 1457–1465, 1999.*

[6] R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, and W.M. Fitch. Predicting the evolution of human influenza A. *Science, Vol 286, pp. 1921–1925, 1999.*

[7] T. Murofushi and M. Sugeno. An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure. *Fuzzy Sets and Systems, Vol 29, pp. 201–227, 1989.*

[8] T. Calvo, A. Kolesarova, M. Komornikova, and R. Mesiar. Aggregation operators: New trends and applications. *Physica-Verlag, Springer, 2002.*

[9] T. Di Matteo, T. Aste, and M.M. Dacorogna. Long term memories of developed and emerging markets: using the scaling analysis to characterize their stage of development. *Journal of Banking & Finance, Vol 29, pp. 827-851, 2005.*

[10] D. Whitford. Proteins: structure and function. *John Wiley & Sons Ltd., 2005.*

[11] T. Murofushi and M. Sugeno. An interpretation of fuzzy measure and the Choquet integral as an integral

with respect to a fuzzy measure. *Fuzzy Sets and Systems, Vol 29, pp. 201–227, 1989.*

[12] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acid Res. Vol 24, pp. 4876–4882, 1997.*

[13] S.D. Black and D.R. Mould. Development of Hydrophobicity Parameters to Analyze Proteins Which Bear Post- or Cotranslational Modifications. *Anal. Biochem., Vol 193, pp. 72–82, 1991.*

[14] H. Tahani and J. Keller. Information Fusion in Computer Vision Using the Fuzzy Integral. *IEEE Transactions on Systems, Man and Cybernetic, Vol 20, pp. 733-741, 1990.*

[15] Z. Wang and G.J. Klir. Fuzzy Measure Theory. *Plenum Press, New York, 1991.*

[16] Hsiang-Chuan Liu, Chin-Chun Chen, Der-Bang Wu, and Yu-Du Jheng. A new weighting method for detecting outliers in IPA based on Choquet integral. *IEEE International conference on Industrial Engineering and Engineering Management 2007, Singapore, Dec. 2~5 2007.*

[17] Hsiang-Chuan Liu. The Choquet integral regression model based on r-complete measure. *Journal of educational research and development, Vol 2, pp. 87-107, 2006. (in Chinese)*

[18] H.E. Hurst, R. Black, Y.M. Sinaika. Long-Term Storage in Reservoirs: An experimental Study. *Constable, London, 1965.*

[19] R. Kalden and S. Ibrahim. Searching for Self-Similarity in GPRS. *PAM, pp. 83-92, 2004.*

出 席 會 議 報 告 表

會議名稱	International Conference on Machine Learning and Cybernetics	會議時間	2009 年 7 月 12 日 12 時 00 分
會議地點	中國保定		
會 議 主 持 人		出 席 人 員	張培均
會議內容	這次的會議是以共同作者的身份前往，發表的論文是模糊測度在蛋白質耐熱性分類上的應用，在 Machine Learning 的領域裡，我們提出利用模糊積分來整合四種蛋白質特徵，來預測蛋白質的耐熱性。關於模糊測度的估計方式，本篇論文提出了四種，未來將試著應用在癌症相關生化路徑網路交互作用上。 與會過程，獲益良多。		

A NOVEL PREDICTING ALGORITHM OF THERMOSTABLE PROTEINS BASED ON CHOQUET INTEGRAL WITH RESPECT TO L-MEASURE AND HURST EXPONENT

JIUNN-I SHIEH [1],YU-LUNG LIU [2], KUEI-JEN LEE [3],PEI-CHUN CHANG [3],YI-CHENG LIU [3]

[1]Department of Information Science and Applications, Asia University, Taiwan
[2]Department of Computer Science and Information Engineering, Asia University, Taiwan
[3]Department of Bioinformatics, Asia University, Taiwan
E-MAIL: jishieh@yahoo.com.tw, liu720402@hotmail.com, pcchang@ asia.edu.tw, kjlee@asia.edu.tw,vicy912@gmail.com

Abstract:

Establishing a good algorithm for predicting temperature of thermostable proteins is an important issue. In this study, a novel thermostable proteins prediction method using Hurst exponent and Choquet integral regression model based on L-measure and γ-support is proposed. The main idea of this method is to integrate the physicochemical properties, fractal property and Choquet integral regression model for amino symbolic sequences with different lengths. For evaluating the performance of this new algorithm, a 5-fold Cross-Validation MSE is performed. Experimental result shows that this new prediction scheme is better than the Choquet integral regression model based on λ-measure and P-measure, respectively and two methods based on Hurst exponent and the traditional prediction models, ridge regression and multiple regression model, respectively.

Keywords:

L-measure; P-measure; λ-measure; Singleton measures; Hurst exponent

6. Introduction

Many experiments and chemical reactions have to be performed in high temperature environment in many fields such as medical industry, foodstuff industry, and cosmetics industry. Furthermore, many materials employ the thermostable proteins as its basic component. These make the research about thermostable proteins a new and important field nowadays. One important issue about thermostable proteins is to predict the temperature of thermostable proteins. In this paper, a prediction algorithm of thermostable proteins by using Hurst exponent and Choquet integral regression model with λ-measure and γ-support is proposed. The contribution of this method is to integrate the physicochemical properties, fractal property and Choquet integral regression model based on our proposed fuzzy measure, L-masure, for amino symbolic sequences with different lengths.

Basically, the procedure of the proposed method is as follows. A thermostable proteins data set was downloaded from the Protein Data Bank (PDB), http://www.rcsb.org, [1]. By substituting four physicochemical quantities of each residue of amino acid in sequence of the thermostable proteins using the four feature scaling estimators, we can obtain four non-symbolic sequences of the thermostable proteins. Then we compute the Hurst exponents of each non-symbolic sequences of the thermostable proteins, so that we can obtained four features of Hurst exponents in each sequences of the thermostable protein. With these extracted features, the Choquet integral regression model based on our proposed fuzzy measure, L-measure, is used to predict the temperature of the 40 thermostable proteins.

For evaluating the performance of this new proposed scheme, the thermostable proteins data experiment is conducted to compare the 5-fold MSE of the proposing methods and those methods based on Hurst exponent and the traditional classification model, multiple regression model.

7. Four scaling estimators of physicochemical properties for each amino acid

Four physicochemical scaling estimators, solvent-accessible surface area, solvent accessibility percentage, electrostatic interaction, and contact energy in the situation of single amino acid were described as follows.

7.1. Solvent accessible surface area (ASA)

The ASA of a protein was obtained using POPS [2], [3] on the web side (mathbio.nimr.mrc.ac.uk/~ffranca/ POPS/),

selecting output residue areas (POPS_R). Both the polar (hydrophilic) and apolar (hydrophobic) surface areas can be obtained from the output residue areas, which were then changed to the percentage of apolar area for each residue in a protein.

7.2. Solvent accessibility percentages

The solvent accessibility percentages of the residues were obtained using the ASAView [4] data base (www.netasa.org/asaview/). Residues were classified to be buried in a protein core as the values between 0-50%, and those were considered to be exposed to solvent when the percentage exceeded 50%.

7.3. Electrostatic interactions

The number of ion pairs (electrostatic interactions) was calculated according to the following criterion [4]: two oppositely charged residues were considered an ion pair if the distance between the oppositely charged atoms of these residues was less than 18Å. Asp, Glu, Arg, Lys and His residues were used to calculate the ion pairs.

7.4. Contact energies

A 20×20 matrix of effective contact energies, the interaction energies between all amino acids pairs, was developed by Miyazawa and Jernigan [5], [6], which was also called MJ matrix. The MJ effective energy (e_{ij}), which is the element of MJ matrix, was derived from all the possible interaction energies, including hydrophobic and solvation energies. Furthermore, the hydrophobic interaction is the dominant contribution to the MJ effective energy. The e_{ij} can be presented as the following equation

$$e_{ij} = e'_{ij} + \frac{e_{ii} + e_{jj}}{2} \tag{1}$$

The e'_{ij} is the mixing term, which is the free energy change upon the mixing of residues of type i and residues of type j when the contacts in self-pairs i-i and j-j are separated to form i-j pairs. The e_{ii} or e_{jj} is the free energy change after the desolvation of residue i or of residue j to form the self-pairs i-i or j-j. The values of e_{ii} or e_{jj} should have high correlation with the hydrophobicity of residue type i or residue type j [5], [6].

The average contact energy of each type of amino acid, e_i, was used in this work, and it is defined as: [5], [6].

$$e_i = \frac{\sum_{j=1}^{20} e_{ij} N_{ij}}{N_{ir}} \tag{2}$$

where

$$N_{ij} = \sum_{p} n_{ij;p} \tag{3}$$

and

$$n_{ir} = \sum_{j \neq 0} n_{ij} \tag{4}$$

The supscript p denotes the total number of contacts in all proteins, n_{ij} is the total number of contacts between i and j types of amino acid residues, and n_{ir} is the total number contacts made by residue type i.

8. Hurst exponent

The Hurst exponent (H) is a statistical measure used to classify time series for long term memory and predictability [7].

In this study, R/S method is used for the estimation of the Hurst exponent: R/S method [8],[9] is based on empirical observations by Hurst in 1965 and estimates H are based on the R/S statistic. It indicates (asymptotically) second-order self-similarity. H is roughly estimated through the slope of the linear line in a log-log plot, depicting the R/S statistics over the number of points of the aggregated series. That is, given a time sequence of observations, w_t define the Series

$$W(t,\tau) = \sum_{u=1}^{t} (w_u - \bar{w}_\tau), 1 \le t \le \tau \tag{5}$$

where

$$\bar{w}_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} w_t \tag{6}$$

Define

$$R(\tau) = \max_{t=1}^{\tau} W(t,\tau) - \min_{t=1}^{\tau} W(t,\tau) \tag{7}$$

and

$$S(\tau) = \sqrt{\left(\frac{1}{\tau} \sum_{t=1}^{\tau} (w_t - \bar{w}_\tau)^2 \right)} \tag{8}$$

In plotting $\log \frac{R(\tau)}{S(\tau)}$ against $\log \tau$, we expect to get a line whose slope determines the Hurst exponent.

9. The multiple linear regression, ridge regression

Let $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim \mathrm{N}(\underline{0}, \sigma^2 I_n)$ be a multiple linear model, $\hat{\underline{\beta}} = (X'X)^{-1} X'Y$ be the estimated regression

coefficient vector, and $\hat{\underline{\beta}}_k = \left(X'X + kI_n\right)^{-1} X'Y$ be the estimated ridge regression coefficient vector, Kenard and Baldwin [1] suggested

$$\hat{k} = \frac{n\hat{\sigma}^2}{\hat{\underline{\beta}}'\hat{\underline{\beta}}} \tag{9}$$

10. Fuzzy measures

10.1. Definition of fuzzy measures [11]

A fuzzy measure μ on a finite set X is a set function $\mu: 2^X \to [0,1]$ satisfying the following axioms:

(i) $\mu(\phi) = 0$, $\mu(X) = 1$ (boundary conditions) (10)

(ii) $A \subseteq B \Rightarrow \mu(A) \le \mu(B)$ (monotonicity) (11)

10.2. Singleton measures [11]

A singleton measure of a fuzzy measure μ on a finite set X is a function $s: X \to [0,1]$ satisfying:

$$s(x) = \mu(\{x\}),\ x \in X \tag{12}$$

$s(x)$ is called the density of singleton x.

10.3. λ-fuzzy measure [11]

For given singleton measures s, a λ-fuzzy measure, g_λ, is a fuzzy measure on a finite set X, satisfying:

(i) $A, B \in 2^X$, $A \cap B = \phi$, $A \cup B \ne X$

$$\Rightarrow g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda g_\lambda(A) g_\lambda(B) \tag{13}$$

(ii) $\prod_{i=1}^{n}\left[1 + \lambda s(x_i)\right] = \lambda + 1 > 0,\ s(x_i) = g_\lambda(\{x_i\})$ (14)

10.4. P-measure [13]

For given singleton measures s, a P-measure, g_P, is a fuzzy measure on a finite set X, satisfying:

$$\forall A \in 2^X \Rightarrow g_P(A) = \max_{x \in A} s(x) = \max_{x \in A} g_P(\{x\}) \tag{15}$$

10.5. L-measure [6], [9]

For given singleton measure s, a L-measure, g_L, is a fuzzy measure on a finite set X, $|X| = n$, satisfying:

(i) $L \in [0, \infty)$ (16)

(ii) $\forall A \subset X,\ n - |A| + (|A| - 1)L > 0 \Rightarrow$

$$g_L(A) = \max_{x \in A}\left[s(x)\right] + \frac{(|A| - 1)L \sum_{x \in A} s(x)\left[1 - \max_{x \in A}\left[s(x)\right]\right]}{\left[n - |A| + (|A| - 1)L\right] \sum_{x \in X} s(x)} \tag{17}$$

11. Choquet integral regression models

11.1. Choquet integral [10]

Let μ be a fuzzy measure on a finite set X. The Choquet integral of $f_i: X \to R_+$ with respect to μ for individual i is denoted by

$$\int_C f_i d\mu = \sum_{j=1}^{n}\left[f_i\left(x_{(j)}\right) - f_i\left(x_{(j-1)}\right)\right]\mu\left(A_{(j)}^i\right), i = 1, 2, ..., N \tag{18}$$

where $f_i\left(x_{(0)}\right) = 0$, $f_i\left(x_{(j)}\right)$ indicates that the indices have been permuted so that

$$0 \le f_i\left(x_{(1)}\right) \le f_i\left(x_{(2)}\right) \le ... \le f_i\left(x_{(n)}\right) \tag{19}$$

$$A_{(j)} = \left\{x_{(j)}, x_{(j+1)}, ..., x_{(n)}\right\} \tag{20}$$

11.2. Choquet integral regression models [12]

Let $y_1, y_2, ..., y_N$ be global or response evaluations of N sample points and $f_1(x_j), f_2(x_j), ..., f_N(x_j)$, $j = 1, 2, ..., n$, be their evaluations of independent variabe x_j, where $f_i: X \to R_+$, $i = 1, 2, ..., N$.

Let μ be a fuzzy measure with γ-support, $\alpha, \beta \in R$,

$$y_i = \alpha + \beta \int_C f_i dg_\mu + e_i,\ e_i \sim N(0, \sigma^2),\ i = 1, 2, ..., N \tag{21}$$

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha, \beta}\left[\sum_{i=1}^{N}\left(y_i - \alpha - \beta \int_C f_i dg_\mu\right)^2\right] \tag{22}$$

then $\hat{y}_i = \hat{\alpha} + \hat{\beta} \int f_i dg_\mu,\ i = 1, 2, ..., N$ is called the Choquet integral regression equation of μ with γ-support, where

$$\hat{\beta} = S_{yf} / S_{ff} \tag{23}$$

$$S_{yf} = \frac{\sum_{i=1}^{N}\left[y_i - \frac{1}{N}\sum_{i=1}^{N} y_i\right]\left[\int f_i dg_\mu - \frac{1}{N}\sum_{k=1}^{N}\int f_k dg_\mu\right]}{N - 1} \tag{24}$$

$$S_{ff} = \frac{\sum_{i=1}^{N}\left[\int f_i dg_\mu - \frac{1}{N}\sum_{k=1}^{N}\int f_k dg_\mu\right]^2}{N - 1} \tag{25}$$

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{\beta}\frac{1}{N}\sum_{i=1}^{N}\int f_i dg_\mu \qquad (26)$$

12. Experiment and result

A thermostable proteins data set is provided by the Protein Data Bank (PDB), http://www.rcsb.org/. There are 40 instances with two classes. Substituting four physicochemical quantities, solvent-accessible surface area, solvent accessibility percentage, electrostatic interaction, and contact energy, for each residue of amino acid in sequence of the thermostable proteins by using the four feature scaling estimators, we can obtained four non-symbolic sequences of the thermostable proteins. Computing the Hurst exponents of each non-symbolic sequences of the thermostable proteins, we can obtain four features represented as Hurst exponents respectively in each sequences of the thermostable protein. The transformed data is listed in Table 2.

The generated data with four features in terms of Hurst exponents is applied to evaluate the leave-one-out classification accuracies of three classifiers: the Choquet integral regression, SVM, and the multiple linear regression by using 5-fold Cross-Validation MSE.

The experimental results are listed in Table 1. We can find that the Choquet integral regression model based on L-measure has the best performance.

Table 1 : MSE of prediction models

HE-Prediction models	5-fold CV MSE
HE-Choquet Integral Regression model with L-measure	14.9257
HE-Choquet Integral Regression model with P-measure	15.2290
HE-Choquet Integral Regression model with λ-measure	15.3812
HE- Ridge regression model	16.3311
HE-Multiple linear regression model	19.7055

13. Conclusions and future works

In this paper, a novel classification algorithm of thermostable proteins combining four feature scaling estimators, Hurst exponent, and the Choquet integral regression model is proposed. For evaluating the performance of this new algorithm, a thermostable proteins data set by using leave-one-out classification accuracy is conducted. Experimental result shows that this new prediction algorithm is useful and better than the traditional two prediction models.

In future, we will consider looking for some improving prediction algorithm of thermostable proteins by using Hurst exponent and other prediction models.

14. Acknowledgements

This paper is partially supported by the National Science Council grant (NSC 97-2410--H-468-014).

References

[20] Protein Data Bank (PDB), http://www.rcsb.org/.

[21] Franca Fealernali and Luigi Cavallo, "Parameter optimized surfaces (POPS): analysis of key interactions and conformational changes in the ribosome", Nucleic Acids Reasearch, 2002, Vol. 30, No 13, pp. 2950-2960, Oxford University press.

[22] Luigi Cavallo, Jens Kleinjung and Franca Fraternali, "POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level", Nucleic Acids Reasearch, 2003, Vol. 31. No.13, pp. 3364-3366. DOI:10.1093/nar/gkg601.

[23] S. Ahmad, M. Gromiha, H. Fawareh, "A. Sarai, ASAView: Database and tool for solvent accessibility representation in proteins", BMC bioinformatics, 2004, Vol. 5, pp. 51-55.

[24] A. Szilagyi, and P. Zavodszky, "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey", Structure Folding and Design, 2000, Vol. 8, 493-504.

[25] S. Miyazawa, R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation", Macromolecules, 1985, Vol. 18, pp. 534-552.

[26] H. E. Hurst: "Long term storage capacity of reservoirs", Transactions of the American Society of Civil Engineers, 1951, Vol.116, pp. 770-799.

[27] H. E. Hurst, R. Black, and Y. M. Sinaika, "Long term storage capacity of reservoirs", An experimental study Constable, London, 1965.

[28] Hsiang-Chuan Liu, Horng-jinh Chang, Kuei-Jen Lee, Jiunn-I Shieh, Wen-Chun Huang, Shin-Ming Huang "A Novel Classification Algorithm of Thermostable Proteins by Using Hurst Exponent and SVM Classifier", Proceedings of the 4th International Conferences on Natural Computation, Jinan,

Shandong, China. . 18-20 October, 2008, Page(s): 24-28. DOI. 2008.852.
[29] M. Sugeno, "Theory of fuzzy integrals and its applications", unpublished doctoral dissertation, Tokyo Institute of Technology, Tokyo, Japan, 1974.
[30] G. Choquet, "Theory of capacities," Annales de l'Institut Fourier, 1953, Vol. 5, pp. 131-295.
[31] Hsiang-Chuan Liu, "The Choquet integral regression model based on r-complete measure", Journal of educational research and development, 2006, Vol. 2, No. 4, pp. 87-107 (in Chinese).

Table ecf2 Hurst exponents of four feature scaling of Thermostable Proteins

Code of Proteins	Temperature	ASA	Electrostatic Interactions	Contact Energy	Solvent Accessibility Percentages
2CVZ	75	0.6559	0.6435	0.3978	0.553
1VPD	37	0.5117	0.6349	0.4814	0.372
1HEX	75	0.6212	0.8094	0.6873	0.7599
1CM7	37	0.5975	0.8063	0.6101	0.5528
1J3N	75	0.7124	0.6768	0.5593	0.4956
1E5M	30	0.6132	0.63	0.463	0.5017
1V8I	75	0.7342	0.6436	0.5099	0.5302
1MP2	37	0.6254	0.5171	0.4296	0.4463
1RJW	65	0.5832	0.7464	0.5373	0.5392
2HCY	30	0.6294	0.632	0.4567	0.511
1O17	85	0.6409	0.761	0.4385	0.455
2BPQ	37	0.6281	0.3409	0.4056	0.505
1EP0	65	0.8782	0.6955	0.6463	0.7016
2IXI	37	0.8204	0.5416	0.6022	0.5604
1ULU	75	0.798	0.6098	0.52	0.5792
2PD4	37	0.7795	0.6212	0.4152	0.4787
2FTZ	80	0.7049	0.6165	0.5334	0.6231
1RTR	37	0.775	0.5714	0.5082	0.5775
1R3U	75	0.7789	0.7885	0.6561	0.7263
1GRV	37	0.7884	0.7388	0.541	0.5326
1Y7T	75	0.5093	0.6392	0.7754	0.7338
5MDH	37	0.6355	0.6716	0.6931	0.7487
1O4U	75	0.6884	0.6218	0.3857	0.4246
1QAP	37	0.7866	0.5896	0.4162	0.3935
1KKJ	65	0.7697	0.4715	0.6382	0.6613
1DFO	37	0.606	0.4759	0.7825	0.8338
1B9B	80	0.5785	0.7805	0.5718	0.6376
1MO0	22	0.5165	0.7064	0.4896	0.554
1EFT	71	0.6162	0.6058	0.5075	0.4391
1EFU	37	0.6188	0.5536	0.4947	0.4843
1XGS	100	0.6651	0.6296	0.5594	0.5031
1MAT	37	0.5434	0.5215	0.4436	0.4638
2PRD	72.5	0.7198	0.6668	0.7226	0.6889
1INO	37	0.7965	0.5308	0.6047	0.4988
1AYG	94	0.6404	0.7141	0.5922	0.5253
2PAC	47.3	0.7569	0.6259	0.5012	0.4128
1Jul	85	0.8471	0.4787	0.5739	0.6153
1PII	35	0.7266	0.4996	0.469	0.4957
3MDS	85	0.7288	0.6151	0.5171	0.6288
3SDP	27.5	0.6461	0.7434	0.5171	0.6288