

修士論文

遺伝子ネットワークの構造を加味した 非負値行列因子分解

岡西 孝真

2019 年 3 月 7 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士(工学) 授与の要件として提出した修士論文である。

岡西 孝真

審査委員：

金谷 重彦 教授 (主指導教員)

笠原 正治 教授 (副指導教員)

MD. ALTAF-UL-AMIN 准教授 (副指導教員)

小野 直亮 准教授 (副指導教員)

黄 銘 助教 (副指導教員)

遺伝子ネットワークの構造を加味した

非負値行列因子分解*

岡西 孝真

内容梗概

遺伝子ネットワークは、マイクロアレイや次世代シーケンサから得られた遺伝子の発現データを基に構築され、遺伝子間における相互作用を表している。用いられる遺伝子数は数百から数千にのぼるため、遺伝子ネットワークは大規模かつ複雑で理解が困難ある。そこで、生物学的知見の獲得やモデルの構築を行うために、小規模なネットワークに分割する。しかし、小規模なネットワークへの分割は、遺伝子発現データあるいは遺伝子ネットワークの持つグラフ構造のどちらかの情報にのみ基づいて行われる。そこで本研究では、より良い解釈が行われるために両データの情報を加味した手法を提案した。提案手法は非負値行列因子分解を基盤とし、ネットワーク構造を加味できるよう改良を加えた更新式を用いて、ベイズ情報量基準に基づいた基底数で解析を行った。その結果、得られた特徴的な遺伝子で構成されたネットワークの解釈を行うと、コラーゲン繊維・エネルギー代謝・細胞分裂に関わる機能とその他に大別された。それらのネットワークは生物学的知見に沿った解釈を行うことができるため、通常の非負値行列因子分解よりも優れていることを示した。

キーワード

遺伝子発現, 遺伝子ネットワーク, 非負値行列因子分解, クラスタリング, ベイズ情報量基準, 卵巣がん

*奈良先端科学技術大学院大学 情報科学研究科 修士論文,
2019 年3 月7 日.

Non-negative Matrix Factorization

considering gene network

Takamasa Okanishi

Abstract

In the present work, we constructed gene networks using gene expression data of ovarian cancer patients based on proposed method. Gene networks are built with expression data from microarray or next-generation sequencer, and show interactions among genes. The massive and complex networks are considerably difficult to comprehend because the genes as components are well into thousands. Therefore, they are generally divided into small networks. They are, however, divided by the information of either gene expression data or graph configuration of a gene network. In the study, new method was applied, which used both of the information simultaneously for better evaluations. It is based on non-negative matrix factorization and introduced the updating-rule improved to take account of the construction of networks. The basic numbers for analysis were determined by Bayesian Information Criterion. As a result, the small networks which mainly consist of featured genes could be reasonably distributed into about four functions respectively related to collagen fiber, energy metabolism, cell division, and the other. Thus, the study demonstrated the new method should be more appropriate rather than normal non-negative matrix factorization.

Keywords:

gene expression, gene network, non-negative matrix factorization, clustering, bayesian information criterion, ovarian cancer

*Master's Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, March 7, 2019.

目次

第 1 章	序論.....	1
1.1	研究背景.....	1
1.2	遺伝子発現とその発現解析技術.....	2
1.3	遺伝子ネットワークとその理解.....	3
1.4	研究目的.....	3
1.5	本論文の構成.....	4
第 2 章	非負値行列因子分解.....	5
2.1	次元削減による特徴抽出.....	5
2.2	NMF の基本的性質.....	5
2.3	アルゴリズム.....	8
2.3.1	損失関数.....	8
2.3.2	補助関数法による更新式の導出.....	9
2.3.3	Lagrange の未定乗数法による更新式の導出.....	12
第 3 章	提案手法.....	14
3.1	ラプラシアン行列.....	14
3.1.1	ラプラシアン行列の定義.....	14
3.1.2	ラプラシアン行列の性質.....	16
3.2	提案手法における更新式の導出.....	17
第 4 章	結果.....	21
4.1	解析データ.....	21
4.2	基底数 k の決定方法.....	21
4.3	特徴的な遺伝子の選択.....	23
4.4	遺伝子ネットワークの描画.....	24
第 5 章	考察.....	68
5.1	提案手法における特徴的な遺伝子の選択.....	68
5.2	遺伝子のネットワーク構造を加味した効果.....	68
5.3	各基底におけるクリニカル情報について.....	69
5.4	AIC による基底数の推定.....	70
5.5	λ の値による基底数の変動.....	71
第 6 章	結論.....	74

謝辭.....	75
参考文献.....	76
付録.....	79

図目次

図 1	DNA マイクロアレイ	2
図 2	NMF の幾何学的イメージ	7
図 3	係数行列 V のスパース性	7
図 4	共起成分をまとめる効果（上：共起しない場合，下：共起する場合） ..	8
図 5	無向グラフにおける隣接行列の例	15
図 6	長さ 2 のノード 1 からノード 5 への経路	15
図 7	ラプラシアン行列の例	16
図 8	基底数 k に対する誤差の遷移	21
図 9	基底数 k に対する BIC の遷移	23
図 10	$k=24$ の基底ベクトルにおける要素値のランクオーダー	23
図 11	$k=24$ における遺伝子ネットワーク	24
図 12	基底における階層的クラスタリング	25
図 13	$k=17$ における遺伝子ネットワーク	26
図 14	$k=25$ における遺伝子ネットワーク	27
図 15	$k=32$ における遺伝子ネットワーク	28
図 16	$k=26$ における遺伝子ネットワーク	29
図 17	$k=13$ における遺伝子ネットワーク	30
図 18	$k=29$ における遺伝子ネットワーク	31
図 19	$k=23$ における遺伝子ネットワーク	33
図 20	$k=33$ における遺伝子ネットワーク	34
図 21	$k=19$ における遺伝子ネットワーク	35
図 22	$k=28$ における遺伝子ネットワーク	36
図 23	$k=20$ における遺伝子ネットワーク	38
図 24	$k=12$ における遺伝子ネットワーク	39
図 25	$k=35$ における遺伝子ネットワーク	40
図 26	$k=1$ における遺伝子ネットワーク	41
図 27	$k=8$ における遺伝子ネットワーク	42
図 28	$k=36$ における遺伝子ネットワーク	43
図 29	$k=31$ における遺伝子ネットワーク	44
図 30	$k=39$ における遺伝子ネットワーク	45
図 31	$k=11$ における遺伝子ネットワーク	46

図 32	k=38 における遺伝子ネットワーク	47
図 33	k=34 における遺伝子ネットワーク	48
図 34	k=6 における遺伝子ネットワーク	49
図 35	k=27 における遺伝子ネットワーク	50
図 36	k=3 における遺伝子ネットワーク	51
図 37	k=21 における遺伝子ネットワーク	52
図 38	k=41 における遺伝子ネットワーク	53
図 39	k=30 における遺伝子ネットワーク	54
図 40	k=16 における遺伝子ネットワーク	55
図 41	k=22 における遺伝子ネットワーク	56
図 42	k=37 における遺伝子ネットワーク	57
図 43	k=4 における遺伝子ネットワーク	58
図 44	k=40 における遺伝子ネットワーク	59
図 45	k=2 における遺伝子ネットワーク	60
図 46	k=5 における遺伝子ネットワーク	61
図 47	k=9 における遺伝子ネットワーク	62
図 48	k=14 における遺伝子ネットワーク	63
図 49	k=10 における遺伝子ネットワーク	64
図 50	k=18 における遺伝子ネットワーク	65
図 51	k=7 における遺伝子ネットワーク	66
図 52	k=15 における遺伝子ネットワーク	67
図 53	NMF での解析結果の一例	69
図 54	基底数 k に対する AIC の遷移 (上 : 5.1 式, 下 : 5.2 式)	71
図 55	各 λ による基底数 k に対する誤差の遷移	72
図 56	各 λ による基底数 k に対する BIC の遷移	72

第1章 序論

1.1 研究背景

がんは医学研究・治療において重要な領域の1つである。平成29年度に厚生労働省が公表した人口動態統計では、悪性新生物による死亡者数は一貫して増加傾向にあり、昭和56年以降の第1位となっている。現在においても主な死因別割合で悪性新生物が最も高い割合を占めている[1]。がんは生活習慣やウイルス感染などによっても発生するが、主な要因として遺伝子の異常が挙げられる。通常、正常な細胞であれば、体の周囲の状況に合わせ増殖したり増殖が停止したりする。例えば皮膚の上皮細胞では、外傷があれば細胞が増殖し傷を塞ぎ、治れば増殖を停止する。またウイルス感染や紫外線などで遺伝子が傷つくと、その細胞の自殺プログラム（アポトーシス）によって自身のがん化や周囲の細胞への悪影響を抑えようとする。このような仕組みは細胞の核に存在する遺伝子によって制御されている。

しかし遺伝子に異常を来すと、正常な仕組みが働かず増殖を繰り返すようになる。増殖によって出来たがん腫瘍が周囲の組織を破壊し、悪化が進むと血流やリンパに乗って他の部位に転移することが知られている。遺伝子の異常とは2つに分類することができ、1つ目は細胞増殖を促す遺伝子の異常である。正常な細胞でも増殖を制御する遺伝子は働いているが、がん細胞ではその遺伝子が過剰に働くため増殖を繰り返すようになる。このような遺伝子はがん遺伝子と呼ばれている。2つ目は細胞のがん化を抑制するがん抑制遺伝子の異常である。がん抑制遺伝子は細胞の異常な増殖の抑制や、遺伝子の異常を修復、アポトーシスへの誘導などの働きを持つ。

がん遺伝子はその他の正常な遺伝子の働きも狂わせることも知られており、がん化に直接関係のある遺伝子をドライバー遺伝子、がん化には無関係な遺伝子の異常をパッセンジャー遺伝子と呼ぶ。がん研究ではこのドライバー遺伝子を特定し振る舞いを知ることで、がん疾患のメカニズムを解明し創薬へ貢献している。ドライバー遺伝子の特定やその振る舞いを知るためには遺伝子間の制御関係を同定あるいは推定する必要がある。遺伝子間の相互作用し合う制御関係のことを遺伝子ネットワークと呼び、生物の生理現象を知るための重要な知見を得ることが出来る。現在に至るまで数多くの遺伝子ネットワークが推定され、ドライバー遺伝子の同定や疾患メカニズムの解明が行われている。

1.2 遺伝子発現とその発現解析技術

遺伝子ネットワークは遺伝子の発現データから構築される。遺伝子は細胞核の中に存在する染色体を形成するデオキシリボ核酸（DNA）上にコードされており、特定のタンパク質の設計図となっている。細胞内では必要に応じたタンパク質を形成するために、まず DNA からリボ核酸（RNA）に転写が行われ、プロセッシングと呼ばれる修飾を経てメッセンジャーRNA（mRNA）が形成される。mRNA が翻訳されることにより最終産物としてタンパク質が生合成される。この DNA 上の遺伝子が mRNA を経てタンパク質に翻訳されることで機能することを遺伝子発現と呼ぶ。

また、遺伝子の発現量を測定するために従来から用いられてきた技術がマイクロアレイである。マイクロアレイとは基板上に多数のプローブ（比較的短い核酸など）を配置し、細胞から抽出した全 mRNA など、それぞれ特異的なプローブに結合させることで、その有無や存在量を測定する技術である。基盤に固定する物質の種類によって様々なバリエーションがあるが、その中でも DNA をプローブに用いた DNA マイクロアレイが最も広く利用されている（図 1）。

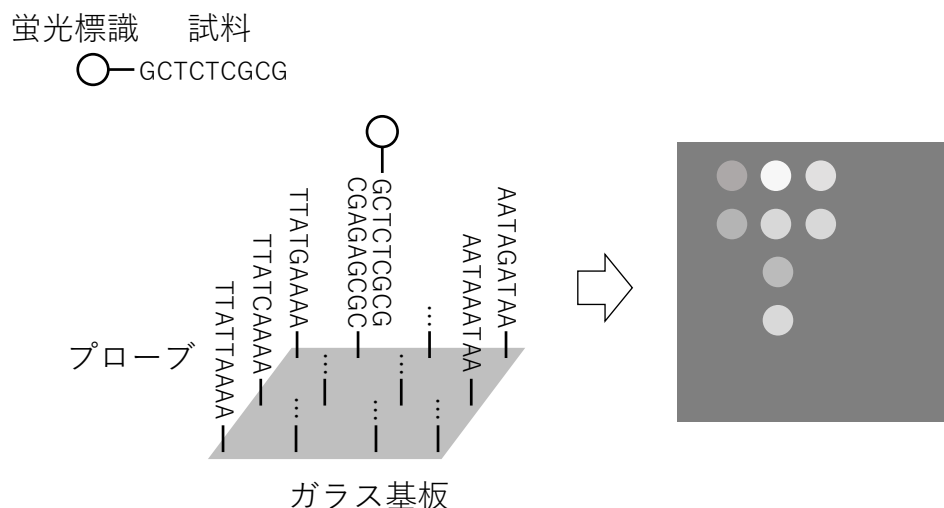


図 1 DNA マイクロアレイ

DNA マイクロアレイ法はその原理上、ゲノムが解読された生物に限定されており、未知の mRNA を検出できない問題があった。この問題点を克服した技術が次世代シーケンサ（NGS）である。NGS は 2005 年に 454 Life Sciences 社が初めて発表されて以降 Illumina

社やライフテクノロジー社などからいくつかの NGS が発表されている。NGS ではエマルジョン PCR やブリッジ PCR と呼ばれる方法を使い、膨大な数の鋳型 DNA がクローニングを介さずに小さなデバイス上に高密度に配置される。これにより扱える生物種の制限が無くなり、ゲノム解読や遺伝子研究を劇的に推し進めることとなった[2].

1.3 遺伝子ネットワークとその理解

マイクロアレイや NGS から得られた発現データを基に遺伝子ネットワークは構築され、遺伝子の相互作用・制御関係はグラフとして表現される。グラフは頂点（ノード）とそれらを結ぶ辺（エッジ）で構成され、遺伝子ネットワークではノード が遺伝子、エッジが制御関係を意味する。またネットワークを構築する手法によりエッジに向きがある有向グラフと、向きのない無向グラフとがある。これまでに様々な遺伝子ネットワークを構築する手法が開発されており、相関ネットワークやベイジアンネットワークなどが存在する[3][4]。それぞれの手法ごとに想定される生物学的仮定や数理モデルが異なっている。例えば、簡便さから広く利用されている相関係数を用いた相関ネットワークでは遺伝子発現パターンが類似していれば生物学的機能も類似しているという仮定が前提であり、エッジの持つ数学的意味合いは相関係数である[5][6]。ベイジアンネットワークが推測するネットワークは非循環であり、フィードバック制御など生物学に見られる循環機構の存在は仮定されておらず（ただし、時系列データを用いた動的ベイジアンネットワークでは循環型のネットワークを得ることができる）、エッジの持つ数学的意味合いは条件付き事後確率である[7][8]。また、推定された遺伝子ネットワーク上で多数の遺伝子と結合している遺伝子はハブ遺伝子と呼ばれ、生物学的に重要な意味を持つという報告がなされており[9][10]、がん研究においてもハブ遺伝子のがん化に直接関わりのあるドライバー遺伝子である可能性を示唆している[11]。

1.4 研究目的

遺伝子ネットワークから遺伝子間の相互作用・振る舞いを知ることにより、生物学的特徴あるいは疾患のメカニズムに関わる知見を得ることができる。ところが遺伝子の数は膨大であり、ヒトの場合 22,000~4,000 個の遺伝子が存在すると言われている。遺伝子ネットワ

ークを構築する際に使用される遺伝子でもその数は約数百～数千個にのぼり、得られる遺伝子ネットワークは大規模且つ複雑なもので、そのままでは理解が困難である。そのため一般的には大規模なネットワークから小規模なネットワークを形成することで、得られた遺伝子ネットワークの解釈が行われる[11][12]。しかし、小規模な遺伝子ネットワークを形成する際、遺伝子発現データの相関あるいは遺伝子ネットワークの持つグラフ構造のどちらか単一の情報のみに基づいて行われるため、全ての情報を加味できていない問題点がある。

そこで本研究では、大規模な遺伝子ネットワークのより良い解釈を行うことを目的とし、遺伝子発現データと遺伝子ネットワークのグラフ情報の両方を加味した手法を提案する。一般的に行われる分割により小規模ネットワークを形成する手法とは異なり、大規模ネットワーク上に特徴部分となる領域のマッピング（クラスタリング）を行う。提案手法は教師なし学習の一種である非負値行列因子分解（Non-negative Matrix Factorization, 以下 NMF）を基盤手法としており、研究目的に合うよう改良し解析を行った。

1.5 本論文の構成

第 2 章では基盤手法となる NMF の原理とアルゴリズムについて述べる。第 3 章では、遺伝子ネットワークを扱う上で生じる NMF の問題点に対応する方法と提案手法について述べ、第 4 章では提案手法による結果を示す。第 5 章では第 4 章の結果についての考察を行い、最後に第 6 章で結論を述べる。

第2章 非負値行列因子分解

この章では機械学習の手法の一種である NMF の基本的な原理等について述べる。NMF は元々画像処理の分野で生まれた技術であり、顔画像から目や鼻などの顔のパーツを抽出するのが目的であった[13]。現在では画像処理の分野だけに留まらず、信号処理・自然言語処理などの分野においても広く活用されている。2.1 節では NMF に関する背景知識、2.2 節では NMF のもつ基本的な性質、2.3 節では NMF のアルゴリズムについてそれぞれ述べる。

2.1 次元削減による特徴抽出

機械学習は得られたデータから知識や法則性を獲得することを目的としており、クラスタリングやモデリングを行うことで未知データの性質や状態遷移を予測することができる。遺伝子発現データなどは高次元であることが常であり、高次元データをそのまま用いた場合、モデルの汎化性能が低いことが知られている。これは使用している特徴量の数（＝次元数）が膨大であるために、ノイズや過学習を引き起こしていることが原因であり、更には計算コストが高くなるデメリットも生じてくる。それらを回避するために次元削減を行う。次元削減とは高次元データから有意義な特徴量を特定することや、新たな軸を作成する（基底の変換する）ことでデータの低次元化を行うことである。次元削減の後、得られた基底はそのデータを成す重要な構成要素として意味解釈が行われる。一般的な手法としては主成分分析（Principal Component Analysis, 以下 PCA）や特異値分解（Singular Value Decomposition, 以下 SVD）、NMF などが存在する。いずれも特徴抽出を行う手法であるが、NMF は他の 2 つとは大きく異なる点を有している。それは得られた基底が全て正の実数で表される点である。PCA や SVD では得られた基底は実数で表され負の値も含むため、解釈が困難になる場合がある。NMF では非負という制約のもと基底の変換を行うため、このような解釈問題が発生しない利点がある。実世界ではパワースペクトルや画素値などといった非負値データが多く、非負値データの構成要素もまた非負値であるというのが NMF の基本概念である。このような利点から本研究では NMF を基盤手法として採用している。

2.2 NMF の基本的性質

NMF のアイデアは 1994 年に Paatero 氏らにより Positive Matrix Factorization として考

案され[14], 1999 年に Lee 氏らによって非負性を保証しながら効率的に解くことのできるアルゴリズムが提案された手法である[13]. NMF は非負値の行列データ $X_{i \times j}$ が与えられた時, 非負値の 2 つの行列 $U_{i \times k}$ と $V_{k \times j}$ による積で近似する手法である.

$$X_{i \times j} \simeq U_{i \times k} V_{k \times j} \quad (2.1)$$

ここで k は, $k < \min(i, j)$ となる基底の数であり, 事前に与えられたパラメータである. NMF はクラスタリング手法の 1 つとして考えることが可能であり, その場合, k がクラスタの数に相当している. NMF では入力データ $X_{i \times j}$ を観測ベクトル $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j \in \mathbb{R}^{\geq 0}$ から成る行列, $U_{i \times k}$ を基底ベクトル $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k \in \mathbb{R}^{\geq 0}$ から成る行列とみなし, 各観測ベクトルは k 個の基底ベクトルの結合係数 $\mathbf{v}_{k1}, \dots, \mathbf{v}_{kj} \in \mathbb{R}^{\geq 0}$ による重みつき和で表され则认为. NMF を適用する際, 加法性が成り立つと仮定していることに注意が必要である. このことから(2.1)式は次のように表すこともできる.

$$\mathbf{X}_j \simeq \sum_{k=1} \mathbf{u}_k \mathbf{v}_{k,j} \quad (2.2)$$

$k < \min(i, j)$ の時, NMF は観測行列 X を低ランク行列で近似することに相当しており, 求められた基底行列や係数行列は重要な意味を持つ. PCA や SVD では観測データが属する部分空間を見つけるのに対し, NMF では観測データに最も良く当てはまる凸錐を見つけるものと解釈できる (図 2).

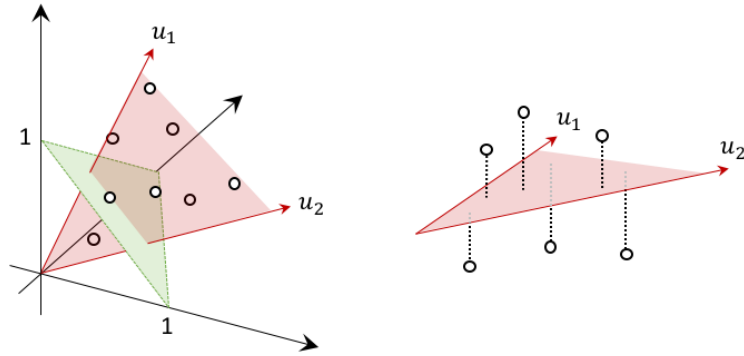


図 2 NMF の幾何学的イメージ

NMF では非負という制約により，係数行列 V に含まれる 1 つ，あるいは少数の要素が値を持ち，それ以外は 0 であるベクトルがいくつか得られる傾向がある．得られる行列 U ， V は観測行列 X との誤差を最小とするように求められるが，この時の乖離距離 $D(X|UV)$ を用いて表すと，図 2 中において各点から凸錐平面に下ろした垂線の総和に相当する．基底 u_1 ， u_2 によって張られる凸錐平面よりも外側に観測点があると係数が負となるので認められず，各基底は全ての観測点が平面上に乗るよう広がるように配置される（図 3）．これによりいくつかの観測点は凸錐平面の境界線上の点となる．これは係数ベクトルの要素のうちいくつかは 0 となることを示しており，NMF の結果にスパース性を持たせる場合，行列 V に対する制約条件が加えられることがある．

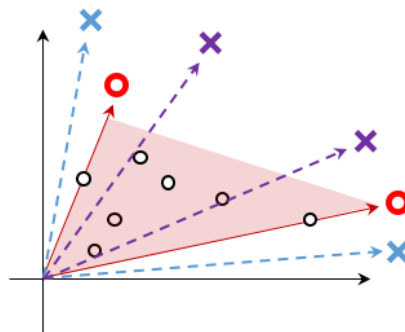


図 3 係数行列 V のスパース性

また，NMF では観測データにおいて共起する成分をまとめた基底ベクトルを得る傾向が

ある．例えば，データが成分 a , b , c の重みつき和で表される時，基底を a , b , c とすればデータを表現することができる．ここでもし， a や b が生起する時常に c も揃って生起する場合，基底を $a + c$ および $b + c$ としてもデータを表現可能である．このように共起する成分同士を基底と置いた方が節約でき，節約した基底分を他の成分を表現するために充てがうことができる（図 4）．NMF では出来るだけ少ない基底でデータを表現することが求められるため，上記のような性質を得られる傾向がある [15]．

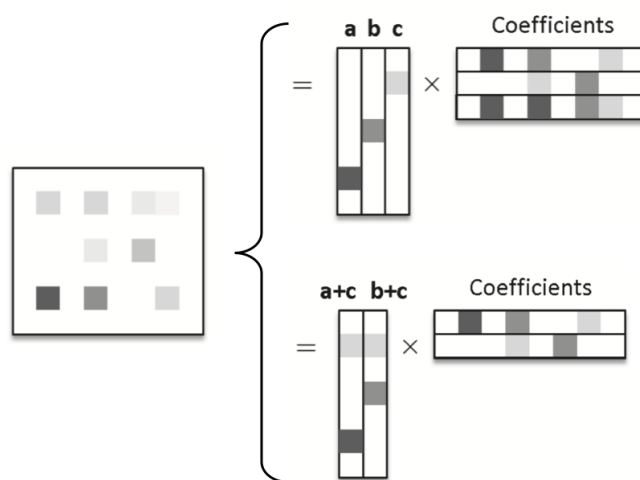


図 4 共起成分をまとめる効果（上：共起しない場合，下：共起する場合）

2.3 アルゴリズム

NMF では入力 X と分解後の UV との誤差をなるべく小さくするように学習を行う．この時の誤差は損失関数として表され，学習アルゴリズムは各行列における各要素の更新式として表現される．この節では具体的な更新式の導出を行う．更新式は異なる 2 つの手法によって導出することができるが，同じ更新式を導くことができる．以下に 2 つの手法による更新式の導出を記す．

2.3.1 損失関数

(2.1) 式の分解は， X と UV の乖離度を評価する損失関数を非負制約下で最小にすることによって求められる．Lee & Seung (1999) では 2 つの損失関数が検討され，制約付き最適化問題のアルゴリズムを導出している．まずもっとも自然な考え方である損失関数が，次の平方ユークリッド距離である．

$$D_{EU}(X|UV) = \sum_{i,j} \left(x_{i,j} - \sum_k u_{i,k} v_{k,j} \right)^2 \quad (2.3)$$

次に、一般化カルバックライブラーダイバージェンス（以下、KL ダイバージェンス）である。これは X と UV のそれぞれの列の合計が1になるように正規化されたものである。

$$D_{KL}(X|UV) = \sum_{i,j} \left(x_{i,j} \log \frac{x_{i,j}}{(UV)_{i,j}} - x_{i,j} + (UV)_{i,j} \right) \quad (2.4)$$

また、損失関数に正則化項を加えることで様々な制約条件を付加することができ上記の損失関数以外にも数多く提案されている。

2.3.2 補助関数法による更新式の導出

NMF ではこれらの損失関数を最小化する非負の行列 U , V を求める。これは非負制約付き非線形最適化問題であり解析的に解くことはできないが、補助関数法を用いることで局所解を得るための反復アルゴリズムを導くことができる。補助関数法とは、目的関数値が下限となる関数（補助関数と呼ぶ）を設計し、目的関数の代わりに補助関数を反復的に降下させることで間接的に目的関数を降下させる方法である[16]。ここでは最も基本的な平方ユークリッド距離に対する更新式の導出を行う。最小化する目的関数を以下に定義する。

$$\begin{aligned} D_{EU}(X|UV) &= \|X - UV\|_F^2 \\ &= \sum_{i,j} \left(x_{i,j} - \sum_k u_{i,k} v_{k,j} \right)^2 \end{aligned} \quad (2.5)$$

(2.5) 式は各要素における近似値との二乗誤差の和、 F はフロベニウスノルムを定義しており、これを最小化することを目的としている。(2.5) 式を展開すると以下の式となる。

$$\begin{aligned}
& \sum_{i,j} \left(x_{i,j} - \sum_k u_{i,k} v_{k,j} \right)^2 \\
&= \sum_{i,j} \left(|x_{i,j}|^2 - 2x_{i,j} \sum_k u_{i,k} v_{k,j} + \left| \sum_k u_{i,k} v_{k,j} \right|^2 \right) \tag{2.6}
\end{aligned}$$

ここで (2.6) 式の第 3 項において、絶対値の中に総和が入っており微分を用いて解析的に解くことができなくなっている。そこで補助関数法を用いて目的関数を下限とする関数を設計し最小化することで、間接的に目的関数を最小化することを行う。(2.6) 式の第 3 項に対し、Jensen の不等式を適応すると、

$$\left| \sum_k u_{i,k} v_{k,j} \right|^2 \leq \sum_k \frac{u_{i,k}^2 v_{k,j}^2}{\lambda_{k,i,j}} \tag{2.7}$$

が得られる。この時 λ は補助関数における変数である。これにより U , V に関する 2 次関数が得られ、微分を行うことができる。(2.7) 式を (2.6) 式に代入することにより、目的関数を下限とする補助関数を以下のように定義することができる。

$$G = \sum_{i,j} \left(|x_{i,j}|^2 - 2x_{i,j} \sum_k u_{i,k} v_{k,j} + \sum_k \frac{u_{i,k}^2 v_{k,j}^2}{\lambda_{k,i,j}} \right) \tag{2.8}$$

Jensen の不等式における等号成立条件から、

$$\lambda_{k,i,j} = \frac{u_{i,k} v_{k,j}}{\sum_k u_{i,k} v_{k,j}} \tag{2.9}$$

である。次に補助関数 G を U , V についてそれぞれ微分すると、

$$\frac{\partial G}{\partial u_{i,k}} = \sum_j \left(-2x_{i,j}v_{k,j} + 2\frac{u_{i,k}v_{k,j}^2}{\lambda_{k,i,j}} \right) = 0 \quad (2.10)$$

$$\frac{\partial G}{\partial v_{k,j}} = \sum_i \left(-2x_{i,j}u_{i,k} + 2\frac{u_{i,k}^2v_{k,j}}{\lambda_{k,i,j}} \right) = 0 \quad (2.11)$$

が得られ、これらを整理すると以下の式になる．

$$u_{i,k} = \frac{\sum_j x_{i,j}v_{k,j}}{\sum_j v_{k,j}^2/\lambda_{k,i,j}}, \quad v_{k,j} = \frac{\sum_i x_{i,j}u_{i,k}}{\sum_i u_{i,k}^2/\lambda_{k,i,j}} \quad (2.12)$$

(2.12)式に補助変数(2.9)式を代入することにより、行列 U 、 V の更新式が以下のように求めることができる．

$$u_{i,k} \leftarrow u_{i,k} \frac{\sum_j x_{i,j}v_{k,j}}{\sum_j v_{k,j} \sum_k' u_{i,k}' v_{k',j}} \quad (2.13)$$

$$v_{k,j} \leftarrow v_{k,j} \frac{\sum_i x_{i,j}u_{i,k}}{\sum_i u_{i,k} \sum_k' u_{i,k}' v_{k',j}} \quad (2.14)$$

また、プログラムでの実装を考慮し、(2.13) (2.14)式を行列表記で以下のように書くことができる．

$$u_{i,k} \leftarrow u_{i,k} \frac{[XV^T]_{i,k}}{[UVV^T]_{i,k}} \quad (2.15)$$

$$v_{k,j} \leftarrow v_{k,j} \frac{[UX^T]_{k,j}}{[U^T UV]_{k,j}} \quad (2.16)$$

NMF ではこの更新式を用いて目的関数が収束するまで計算を行う．

2.3.3 Lagrange の未定乗数法による更新式の導出

以上のような補助関数法以外にも、Lagrange の未定乗数法を用いて更新式を求めることができる[17][18]。以下より Lagrange の未定乗数法による更新式の導出を記す。また、式簡略化のため行列表記で表す。

Lagrange では、以下のような最適化問題を解くことにあたる。

$$\begin{aligned} & \text{minimize} \quad \|X - UV\|_F^2 \\ & \text{subject to} \quad U \geq 0, V \geq 0 \end{aligned}$$

ここから、Lagrange 未定係数 ψ , φ を導入し以下のように定義できる。

$$\begin{aligned} \mathcal{L} &= \|X - UV\|^2 + \text{tr}(\psi U^T) + \text{tr}(\varphi V^T) \\ &= \text{tr}(XX^T) - 2\text{tr}(X^T UV) + \text{tr}(\psi U^T) + \text{tr}(\varphi V^T) \end{aligned} \quad (2.17)$$

Lagrange 関数 $\mathcal{L}$ をそれぞれ U , V について微分すると以下の式が得られる。

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV^T + 2UVV^T + \psi = 0 \quad (2.18)$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^T U + 2V^T U^T U + \varphi = 0 \quad (2.19)$$

ここで KKT (Karush-Kuhn-Tucker) 条件[20]より、以下の等式が成り立つ。

$$\left(-2XV^T + 2UVV^T\right)_{i,k} u_{i,k} = 0 \quad (2.20)$$

$$\left(-2X^T U + 2V^T U^T U\right)_{k,j} v_{k,j}^T = 0 \quad (2.21)$$

(2.20) (2.21) 式に基づくと、それぞれ (2.15) (2.16) 式と同様の更新式が求まる。

以上のように、NMF では更新式さえ求まれば数行のコードで実装を行うことができる。
ここではユークリッド距離での損失関数を最小化する問題を解いたが、他の損失関数においても同様に導出することができる。

第3章 提案手法

NMF は非負値の行列データであれば入力として受け入れるので、遺伝子発現データにも適用することが可能である。しかし、2.3 節における更新式の導出の際、遺伝子ネットワークのもつグラフの幾何学的構造情報を加味できておらず、遺伝子発現データによる単一の情報しか用いられていない。このことは遺伝子ネットワークを扱う上で、手元にある情報を全て活かしきれていない問題点が挙げられる。

また、NMF の性質上、共起成分を同じ基底にまとめるため、データにおける偽相関の影響をそのまま受ける問題点が挙げられる。これにより、様々な機能をもつ遺伝子が同じ基底に格納される恐れがあり、最終的な基底および大規模ネットワークの解釈が困難になる可能性が考えられる。これらの問題点を解決すべく NMF に遺伝子ネットワークのもつグラフ構造を加味させるためにラプラシアン行列を導入する。

3.1 節では、ラプラシアン行列について述べ、3.2 節では NMF にラプラシアン行列を導入した提案手法のアルゴリズムについて述べる。

3.1 ラプラシアン行列

グラフ構造を行列式で表した際、その行列が様々な特性をもつことによりグラフのもつ性質が行列を通して明らかにすることができる。3.1.1 項では、グラフ理論で用いられる基本的な隣接行列からラプラシアン行列の導出を記す。3.1.2 項ではラプラシアン行列がもつ性質について述べる [19]。

3.1.1 ラプラシアン行列の定義

以下ではまず無向グラフの隣接行列 A について定義する。無向グラフの頂点（ノード）の集合を $V = \{1, 2, \dots, n\}$ 、辺（エッジ）の集合を E とする。ノード i とノード j の間にエッジが存在する場合 $(i, j) \in E$ である。この時、隣接行列 A は以下のように定義される。

$$A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

隣接行列は対称行列であり，自分自身とのエッジは考えないため対角成分は全て 0 である．グラフが図 5 のように与えられた時，次のように表すことができる．

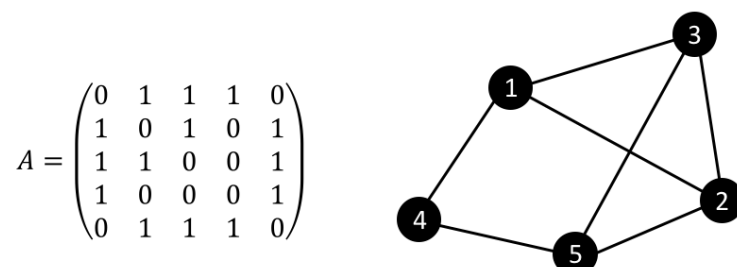


図 5 無向グラフにおける隣接行列の例

また，隣接行列の $A^k (k = 2, 3, 4, \dots)$ の (i, j) 成分はノード i からノード j へ長さが k の経路数を表している．図 5 のグラフに当てはめれば，ノード 1 からノード 5 へ長さ 2 の経路数は次の行列の $(1, 5)$ 成分から知ることができる．

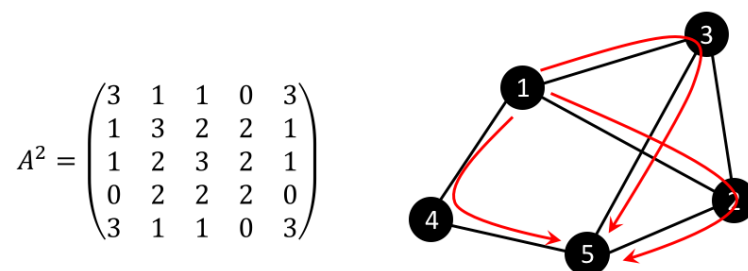


図 6 長さ 2 のノード 1 からノード 5 への経路

このように隣接行列からグラフの性質を代数的に知ることができる．

次にノード i の次数（そのノードがもつエッジの数）を d_i とすると，次数行列 D は以下のよう定義される．

$$D_{ij} = \begin{cases} d_i & (i = j) \\ 0 & (i \neq j) \end{cases} \quad (3.2)$$

以上の隣接行列 A と次数行列 D からラプラシアン行列 L は以下のように定義される．

$$L = D - A \quad (3.3)$$

ラプラシアン行列はグラフラプラシアンとも呼ばれ、隣接行列と同様にグラフの性質を代数的に調べることができる。図 7 はノード数 4 の簡単なグラフからなるラプラシアン行列の例である。

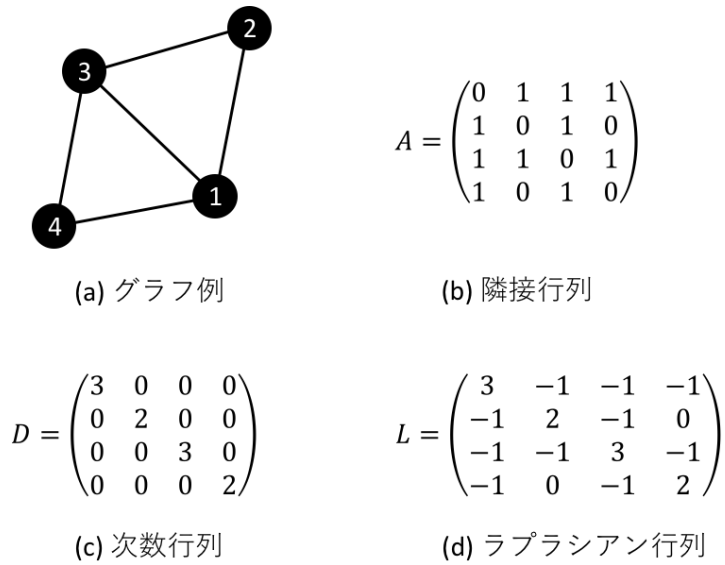


図 7 ラプラシアン行列の例

3.1.2 ラプラシアン行列の性質

ラプラシアン行列には、対角成分がノードのもつエッジ数、それ以外の成分が接続されているノードといったグラフの情報が含まれている。

また、ラプラシアン行列は実対称行列であるため、以下のような性質をもつ。

N 次元のラプラシアン行列の時、

1. 固有値は重複も数えると全て実数で N 個存在する。
2. 固有ベクトルの各要素は全て実数になる。
3. 固有ベクトルが正規直交基底になるよう N 個得られる。

これらの性質より、グラフ理論でのグラフ分割問題において、ラプラシアン行列の固有値問題としてよく解析に用いられる。例えば、ラプラシアン行列の固有値が示すグラフの性質として、0 固有値の数だけグラフにおける離散的な島（連結成分）が存在し、必ず非負であ

ることが知られている．つまりラプラシアン行列は常に 0 固有値が 1 つ存在し，このことはラプラシアン行列の定義から明らかである．ラプラシアン行列の各行を総和すると，

$$3 + (-1) + (-1) + (-1) = 0$$

$$(-1) + 2 + (-1) + 0 = 0$$

$$(-1) + (-1) + 3 + (-1) = 0$$

$$(-1) + 0 + (-1) + 2 = 0$$

となり，0 となることがわかる．これは $\vec{v} = (1, 1, 1, 1)^T$ が L の固有値 0 に属する固有ベクトルであることを示している．

$$\begin{aligned} L\vec{v} &= \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = 0 \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = 0 \times \vec{v} \end{aligned}$$

ラプラシアン行列は以上に挙げたようにグラフの情報を有している．本研究ではラプラシアン行列の固有値問題自体は解かないが，この行列をアルゴリズムに組み込むことによって，NMF が数学的にグラフ情報を加味できるようにしている．

3.2 提案手法における更新式の導出

本研究では，遺伝子ネットワークの構造を加味した NMF を提案する．提案手法では，遺伝子発現データと推定された遺伝子ネットワーク構造のデータの 2 つを使用する．

まず，NMF を以下のように定義する．

$$X_{i \times j} \simeq U_{i \times k} V_{k \times j}^T \quad (3.4)$$

$X_{i \times j}$ は遺伝子発現データであり、 i は遺伝子、 j は患者を意味する．分解後の行列 $U_{i \times k}$ は遺伝子に関する、 $V_{j,k}$ は患者に関する次元削減且つクラスタリングの結果を表す．

次に、行列 U において、遺伝子ネットワーク上で結合をもつ遺伝子間の行ベクトル $\mathbf{u}_m, \mathbf{u}_n$ に重み付けを行った正則化項を定義する．以下に隣接行列 A と正則化項 R を記す．この時、 $m, n \in i$ である．

$$A_{m,n} = \begin{cases} 1 & \text{if gene } m \text{ and } n \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$$

$$\begin{aligned} R &= \frac{1}{2} \sum_{m,n}^i |\mathbf{u}_m - \mathbf{u}_n|^2 A_{m,n} \\ &= \sum_{m=1}^i \mathbf{u}_m^T \mathbf{u}_m D_{m,m} - \sum_{m,n}^i \mathbf{u}_m^T \mathbf{u}_n A_{m,n} \\ &= \text{tr}(U^T D U) - \text{tr}(U^T A U) = \text{tr}(U^T L U) \end{aligned} \quad (3.6)$$

(3.4) (3.6) 式から次のような最適化問題を設定する．

$$\begin{aligned} &\text{minimize} \quad ||X - UV^T||_F^2 + \lambda \text{tr}(U^T L U) \\ &\text{subject to} \quad U \geq 0, V \geq 0 \end{aligned}$$

この時 λ はグラフ情報による影響度を定めるハイパーパラメータであり、本研究では $\lambda = 1$ に設定している．遺伝子間に結合が存在する場合、NMFは誤差を最小化する働きから $\mathbf{u}_m$ と $\mathbf{u}_n$ が近い値に更新されていく．これにより、結合をもつ遺伝子同士は同じ基底ベクトルに格納され、結果的にグラフ構造を加味した基底を得ることができる．

上記の最適化問題を 2.3.3 項における Lagrange の未定乗数法での方法を用いると、Lagrange 関数は、

$$\begin{aligned}
\mathcal{L} &= |X - UV^T|^2 + \lambda \text{tr}(U^T LU) + \text{tr}(\psi U^T) + \text{tr}(\varphi V^T) \\
&= \text{tr}(XX^T) - 2\text{tr}(XVU^T) + \text{tr}(UV^T VU^T) \\
&\quad + \text{tr}(\psi U^T) + \text{tr}(\varphi V^T)
\end{aligned} \tag{3.7}$$

と定義され、それぞれ U , V について微分すると以下の式が得られる.

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^T V + 2\lambda LU + \psi = 0 \tag{3.8}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -2X^T U + 2VU^T U + \varphi = 0 \tag{3.9}$$

ここで KKT 条件[20]から以下の等式が成り立つ.

$$(-2XV + 2UV^T V + 2\lambda LU)_{i,k} u_{i,k} = 0 \tag{3.10}$$

$$(-2X^T U + 2VU^T U)_{j,k} v_{j,k} = 0 \tag{3.11}$$

(3.10) (3.11) 式に基づいて、提案手法における更新式は以下のように求めることができる.

$$u_{i,k} \leftarrow u_{i,k} \frac{[XV + \lambda AU]_{i,k}}{[UV^T V + \lambda DU]_{i,k}} \tag{3.12}$$

$$v_{j,k} \leftarrow v_{j,k} \frac{[X^T U]_{j,k}}{[VU^T U]_{j,k}} \tag{3.13}$$

以上のように、グラフ構造を加味した更新式の導出を行った. 提案手法では大規模な遺伝子ネットワークを、遺伝子発現データまたは遺伝子ネットワーク構造のどちらかだけでは

なく、両方の情報を加味した解析を行うことができる。NMFにおける基底（構成要素）の加算性から、各基底における特徴的な遺伝子ネットワーク（版画における版のようなもの）の重ね合わせとして大規模な遺伝子ネットワークを理解することができ、共起成分をまとめる性質から、それぞれの遺伝子ネットワークにおける共通部分を介した互いの連動性を見ることができる。これらのNMFの利点に加えて複数の情報が合わさることで、結果として大規模な遺伝子ネットワークのより良い解釈を行うことができる。次章では提案手法を用いた解析結果を報告する。

第4章 結果

第3章では提案手法について述べた。第4章では提案手法での解析結果を述べる。まず、解析に用いたデータと基底数 k の決定方法について述べ、その基底で抽出されたネットワークの性質について述べる。

4.1 解析データ

本研究では、TCGA(The Cancer Genome Atlas)に登録されているデータを用いた。TCGAは2006年から米国で開始された大型がんゲノムプロジェクトであり、20種類以上のがん種について、ゲノム・メチル化異常、遺伝子・タンパク質発現異常などの解析データが登録されている。TCGAから卵巣がんの遺伝子発現データのうち、患者の臨床情報などに欠損がない372人分のデータを解析に用いた。

また、卵巣がんの遺伝子発現データのうちTCGAのデータで解析が行われた論文[11]で用いられている245個の遺伝子を選択し、これらの遺伝子間で推定された遺伝子ネットワークの情報を解析に用いた。解析に用いた遺伝子リストは付録にて表に記す。

4.2 基底数 k の決定方法

NMFでは基底の数が多いほど入力データをより細かく再現することができるため、 k の値が大きくなればなるほど分解後の誤差は小さく抑えることができる(図8)。しかし、 k の値が大きすぎると複雑なモデルとなるため、最適な k を決定する必要がある。

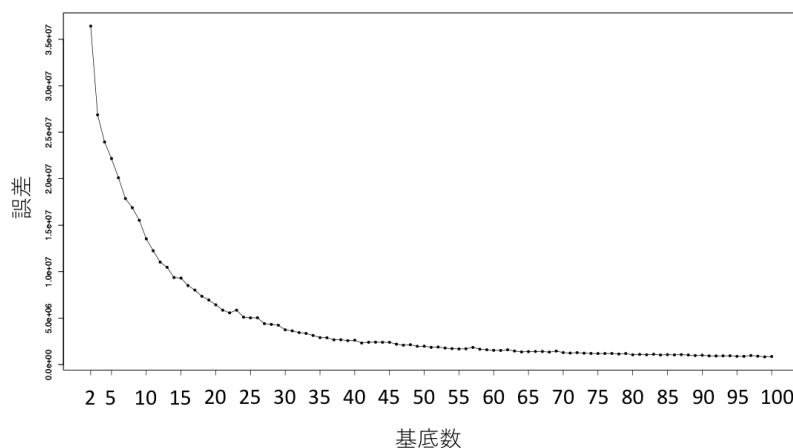


図8 基底数 k に対する誤差の遷移

そこで、ベイズ情報量基準 (Bayesian information criterion 以下, BIC) を用いて最適な k の値を決定する. BIC は数学的に次のように表される.

$$BIC = -2\log(\text{Likelihood}) + k\ln(n) \quad (4.1)$$

ここで, k はモデルラメータの数, n はサンプルサイズである. 第一項はモデルの当てはまりの良さを意味し, サンプルサイズが大きくなれば値は小さくなる. 第二項は罰則項であり, 推定されるモデルが複雑にならないように抑制する. BIC の他にも赤池情報量基準 (Akaike's Information Criterion, 以下 AIC) があるが, AIC と BIC には良いモデルとしての定義が異なっている. AIC では予測精度が優れた (汎化誤差が小さな) モデルが良いモデルとされ, BIC では真のモデルである確率が高い (分布が近い) モデルが良いモデルとしている. NMF での BIC は Owen らによって定義されており [21], これに従って以下のように定義する.

$$BIC = \log(D_{EU}) + k \frac{i+j}{ij} \log \frac{ij}{i+j} \quad (4.2)$$

ここでの D_{EU} は提案手法での最小化目的関数である. (4.2) 式に基づき k の値を変化させた時の BIC の遷移を図 9 に示す. 提案手法では BIC が最小となっている $k=41$ で設定し, 解析を行った.

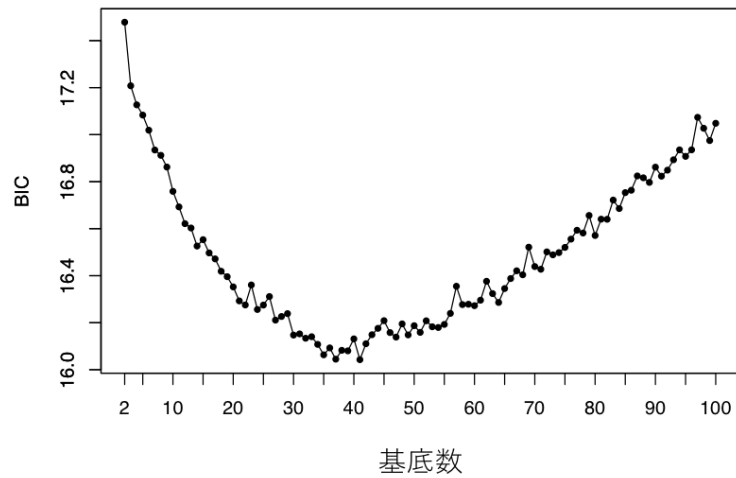


図 9 基底数 k に対する BIC の遷移

4.3 特徴的な遺伝子の選択

次に、解析で得られた 41 個の基底での特徴的な遺伝子群の抽出を行った。抽出された遺伝子間にエッジが存在すれば小規模な遺伝子ネットワークが表現され、大規模ネットワークにおける特徴的な部分であると解釈できる。例として $k=24$ の基底ベクトルについて述べる。行列 U における 24 番目の基底ベクトルに含まれる要素値をランクオーダーで図 10 に示す。

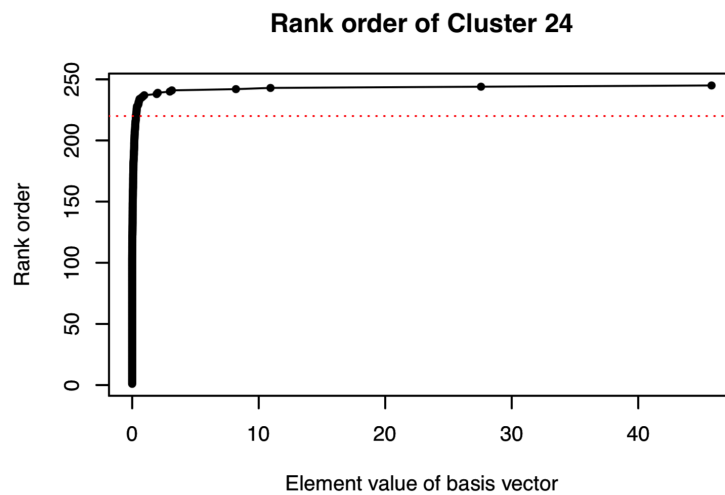


図 10 $k=24$ の基底ベクトルにおける要素値のランクオーダー

図 10 より、基底ベクトルに含まれる要素値のほとんどが 0 であることが分かり、これは

他の基底ベクトルでも同様に見受けられる．そこで要素値が上位 10%（図 10 中の赤点線より上の 25 個）の遺伝子を選択し，その遺伝子間における遺伝子ネットワークを抽出した．次項の図 11 で $k=24$ において抽出された遺伝子ネットワークを示す．図中では要素値の大きさを，円の大きさと色の濃さで表現している．

4.4 遺伝子ネットワークの描画

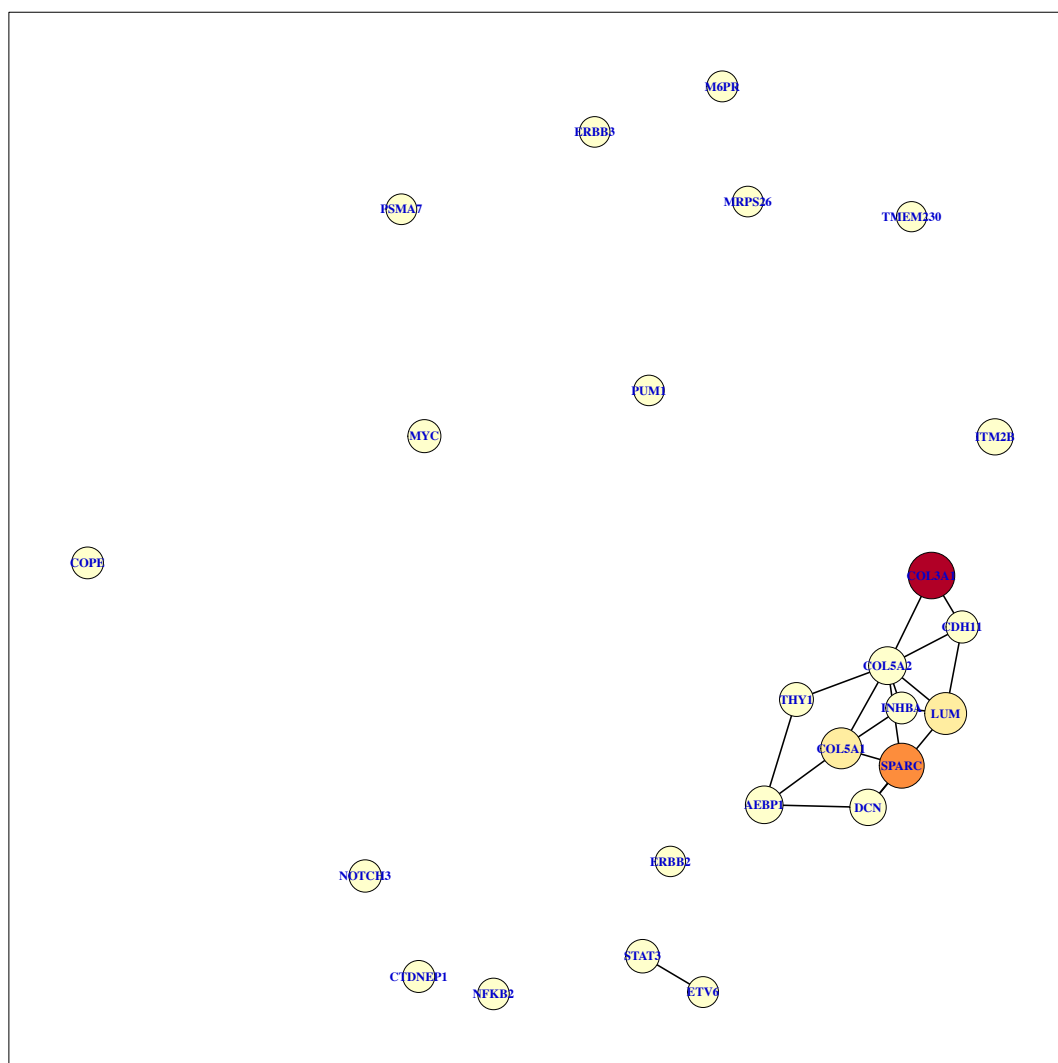


図 11 $k=24$ における遺伝子ネットワーク

抽出された遺伝子ネットワークでは，SPARC, LUM, COL3A1, COL5A1 などを含む遺伝子ネットワークが特徴的に現れている．SPARC はコラーゲン結合マトリックス細胞のタ

ンパク質をコードし、細胞間における相互作用を調節だけでなく、増殖阻害特性から癌抑制遺伝子として知られている。LUM は、コラーゲン原線維の組織化および円周方向の成長・角膜の透明性・上皮細胞の移動および組織修復を担うタンパク質をコードしている。COL3A1 と COL5A2 は同様に、コラーゲン繊維のサブユニットをコードしている。これらのことから、 $k = 24$ で得られた遺伝子ネットワークは、細胞の伸縮系や組織形成に関わるコラーゲン繊維に関する遺伝子ネットワークであると解釈することができる。

次に、41 個の基底ベクトルにおける類似度について階層的クラスタリングを用いてデンドログラムで表現した (図 12)。距離についてはユークリッド距離、併合方法についてはウォード法を用いており、クラスタ数は 4 つに設定した。

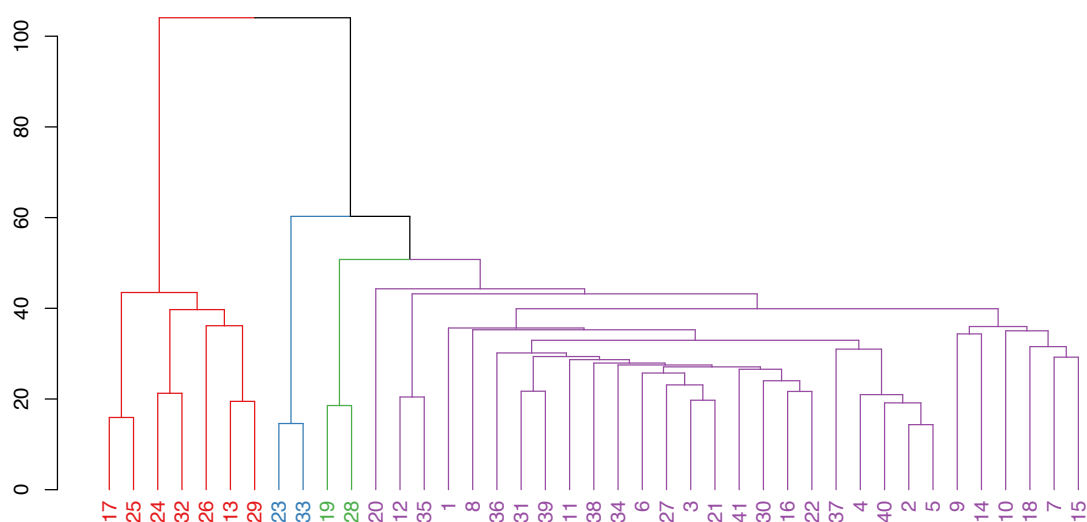


図 12 基底における階層的クラスタリング

便宜上、図 12 の左から順にクラスタ 1、クラスタ 2、クラスタ 3、クラスタ 4 と呼ぶこととする。先ほど例に挙げた $k = 24$ はクラスタ 1 に属し、これと類似した基底群の遺伝子ネットワークを並べた図を以下に示す。

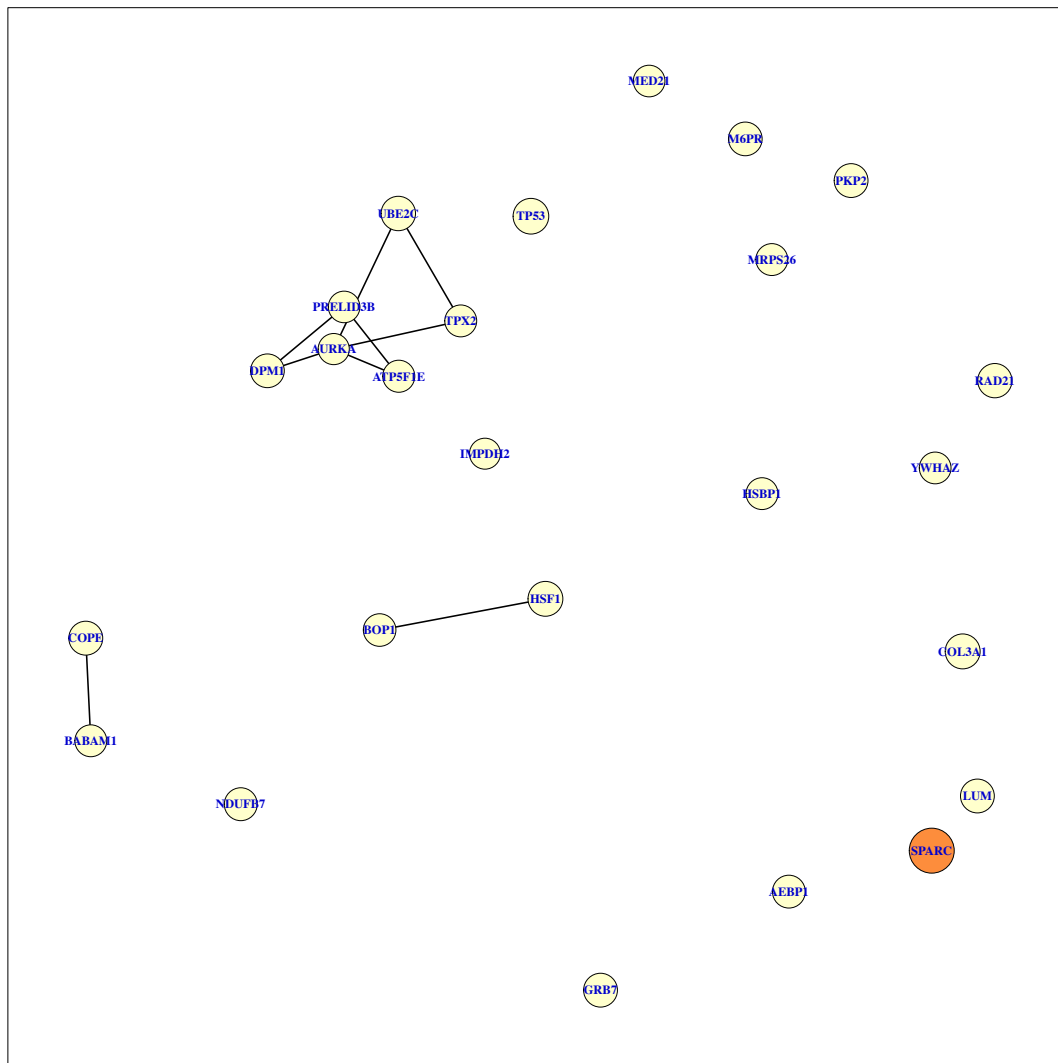


図 13 $k=17$ における遺伝子ネットワーク

$k = 17$ の基底では TPX2, UBE2C, AURKA, ATP5F1E, PRELID3B, DPM1 から成る遺伝子ネットワークがある (図 13). TPX2, UBE2C は細胞周期や有糸分裂に関わる遺伝子であり, AURKA は染色体分裂期の微小管形成に関わる遺伝子として知られている. ATP5F1E はミトコンドリアの ATP 合成に関わる遺伝子である. ATP とはアデノシン三リン酸の英略であり, 細胞内での様々な活動に必要なエネルギーの源である. PRELID3B はヒストンタンパク質をコードしている遺伝子であり, ATP5F1E と同様ミトコンドリアに関係する遺伝子である. DPM1 がコードするたんぱく質は小胞体内に存在し, 翻訳されたたんぱく質の糖鎖修飾に関わっている. これらのことから, コラーゲン繊維と細胞分裂に関わる遺伝子ネットワークであると解釈できる.

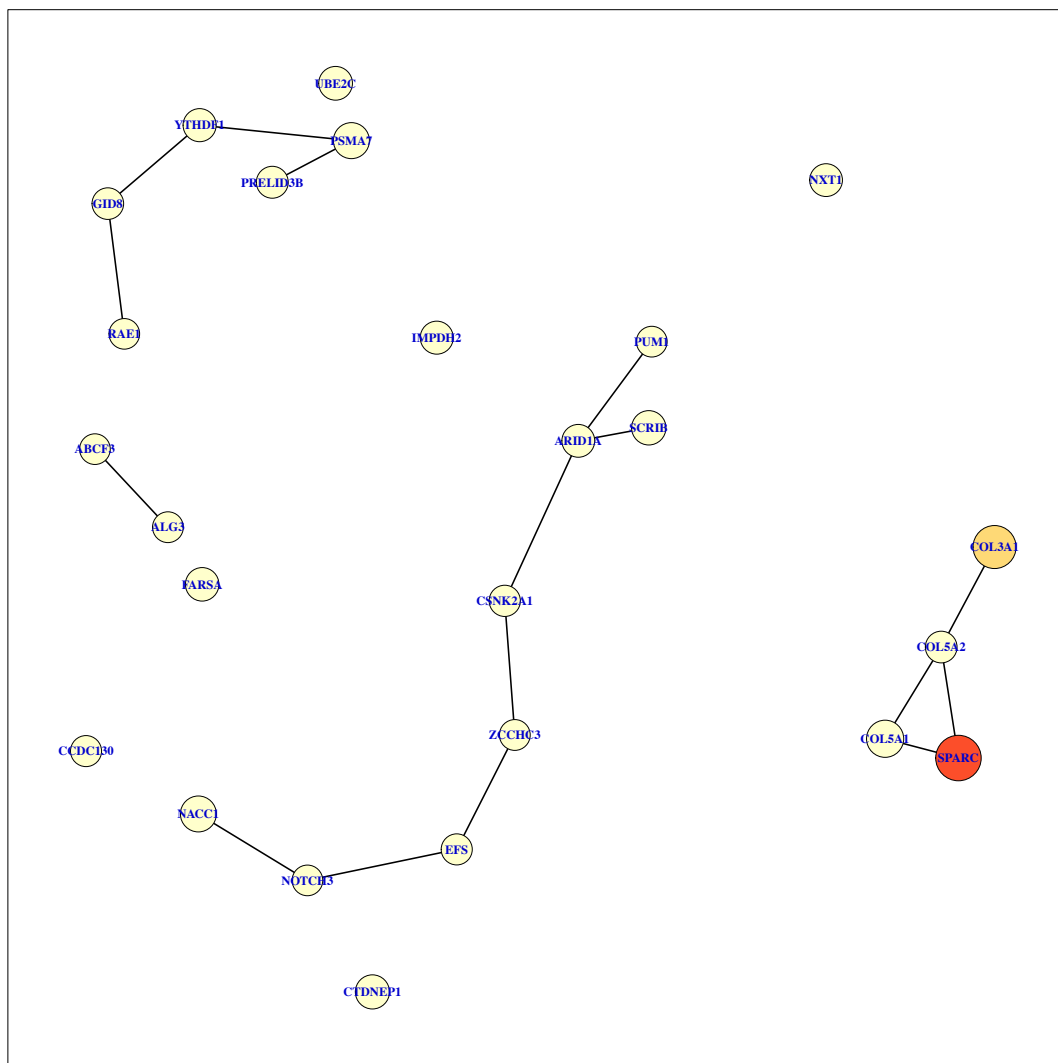


図 14 $k=25$ における遺伝子ネットワーク

$k=25$ の基底では PUM1, SCRIB, ARID1A, CSNK2A1, ZCCHC3, EFS, NOTCH3, NACC1 から成る遺伝子ネットワークがある (図 14). これらは細胞増殖に関わる遺伝子である. その中でも CSNK2A1 と NOTCH3 は発がんに関わる遺伝子であり, ARID1A は腫瘍抑制効果をもつ複合体の形成を促進するがん抑制遺伝子である. これらのことから, コラーゲン繊維とがん細胞における増殖に関わる遺伝子ネットワークであると解釈できる.

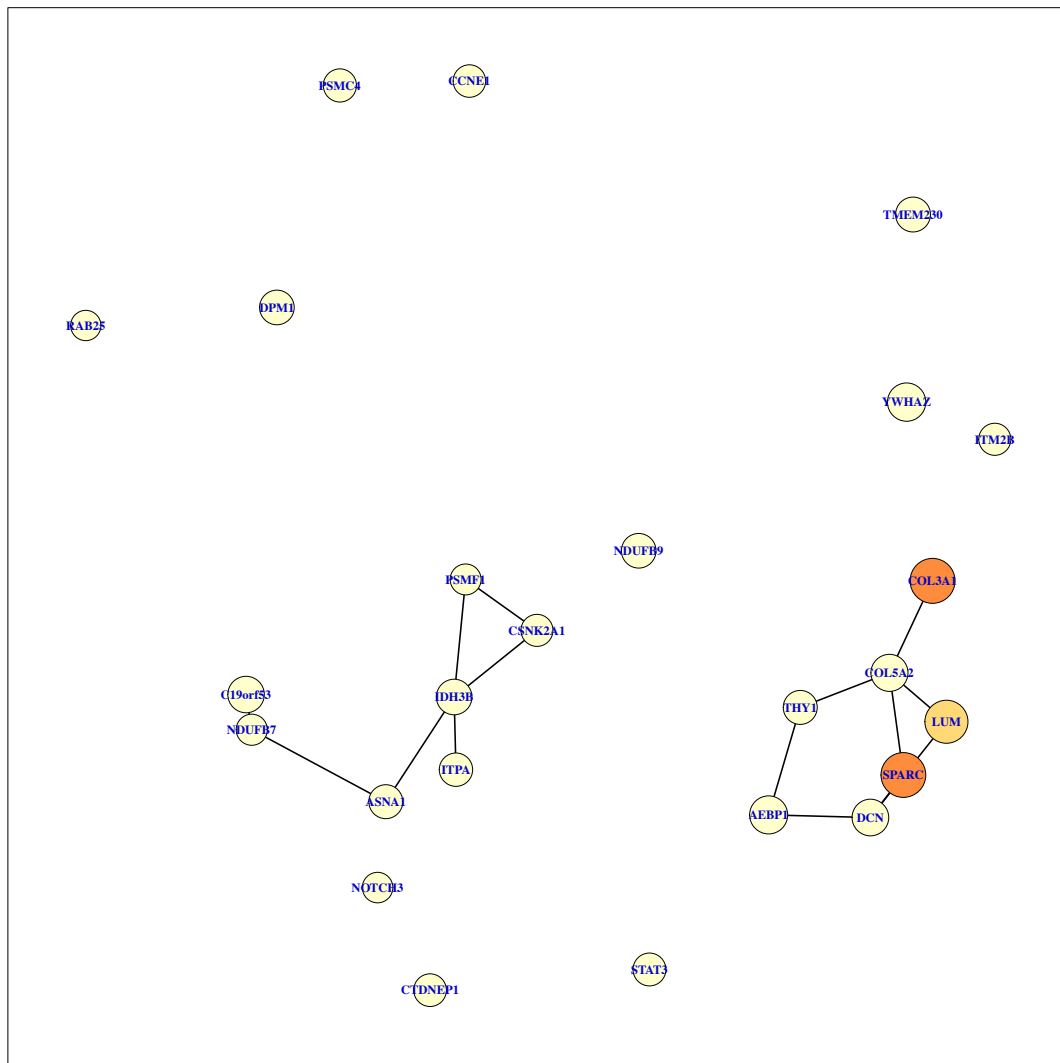


図 15 $k=32$ における遺伝子ネットワーク

$k = 32$ の基底では CSNK2A1, PSMF1, IDH3B, ITPA, ASNA1, NDUF7, C19orf53 から成る遺伝子ネットワークがある (図 15). CSNK2A1 は発がん性タンパク質をリン酸化することでも知られた癌遺伝子であり, PSMF1 と IDH3B はエネルギー源である ATP の触媒や分解に関わる遺伝子である. ASNA1 はタンパク質輸送に関わるタンパク質をコードしている. タンパク質輸送やリン酸化には ATP が必要であることから, 生物学的知見に合った相互作用が得られている. ITPA は結腸や肝臓などの様々ながん組織で過剰に発現していることが知られ, C19orf53 は卵巣がんに関わる癌遺伝子として同定されている. NDUF7 はミトコンドリア内膜に存在する複合体 I のサブユニットをコードしている. これらのことから, コラーゲン繊維と発がんのエネルギーに関わる遺伝子ネットワークであると解釈でき

る.

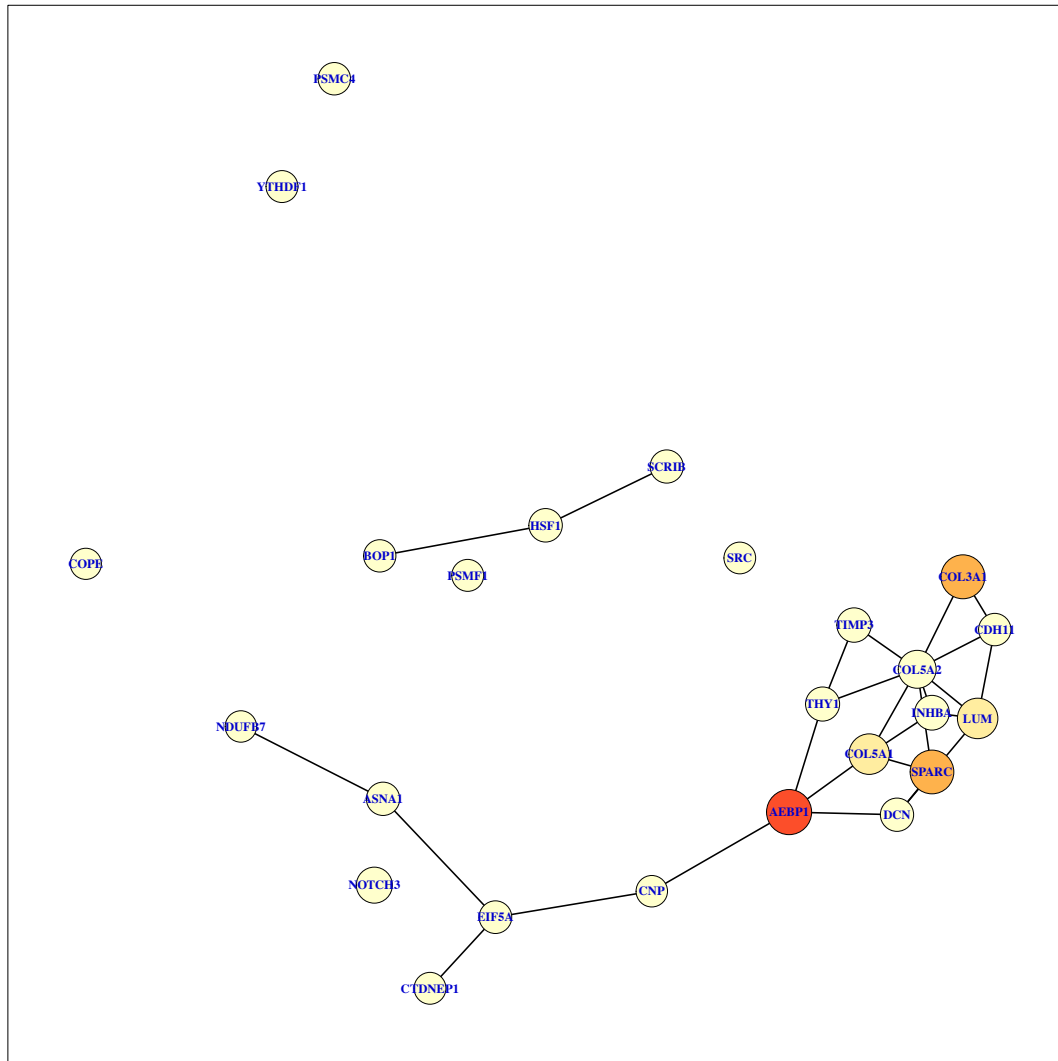


図 16 $k=26$ における遺伝子ネットワーク

$k = 26$ の基底ではコラーゲン繊維に関わる遺伝子ネットワークに続き, AEBP1からCNP, EIF5A, CTDNEP1, ASNA1, NDUFB7 へ連なる遺伝子ネットワークと, BOP1, HSF1, SCRIB から成る遺伝子ネットワークがある (図 16). AEBP1 は転写抑制因子として機能するタンパク質をコードしており, AEBP1 の遺伝子変異により不完全なコラーゲン組織を形成することが知られている. EIF5A と CTDNEP1 は細胞周期に関わるタンパク質をコードしている. ASNA1 と NDUFB7 はそれぞれ小胞体とミトコンドリアに関係するタンパク質をコードしており細胞の様々な機能で相互作用し合っている. 2'.3'-環状ヌ

クレオチド 3'ホスホジエステラーゼの英略である CNP は未だまとまった生理学的な機能について理解されていないが、コラーゲン繊維の調節に関わることも知られており、コラーゲン繊維と細胞周期に関わる遺伝子ネットワークであると解釈できる。加えて、BOP1, HSF1, SCRIB は発がん性やがんの進行に関わるがん遺伝子として知られている。これらのことからコラーゲン繊維と細胞周期、がん化に関わりのある遺伝子ネットワークであると解釈できる。

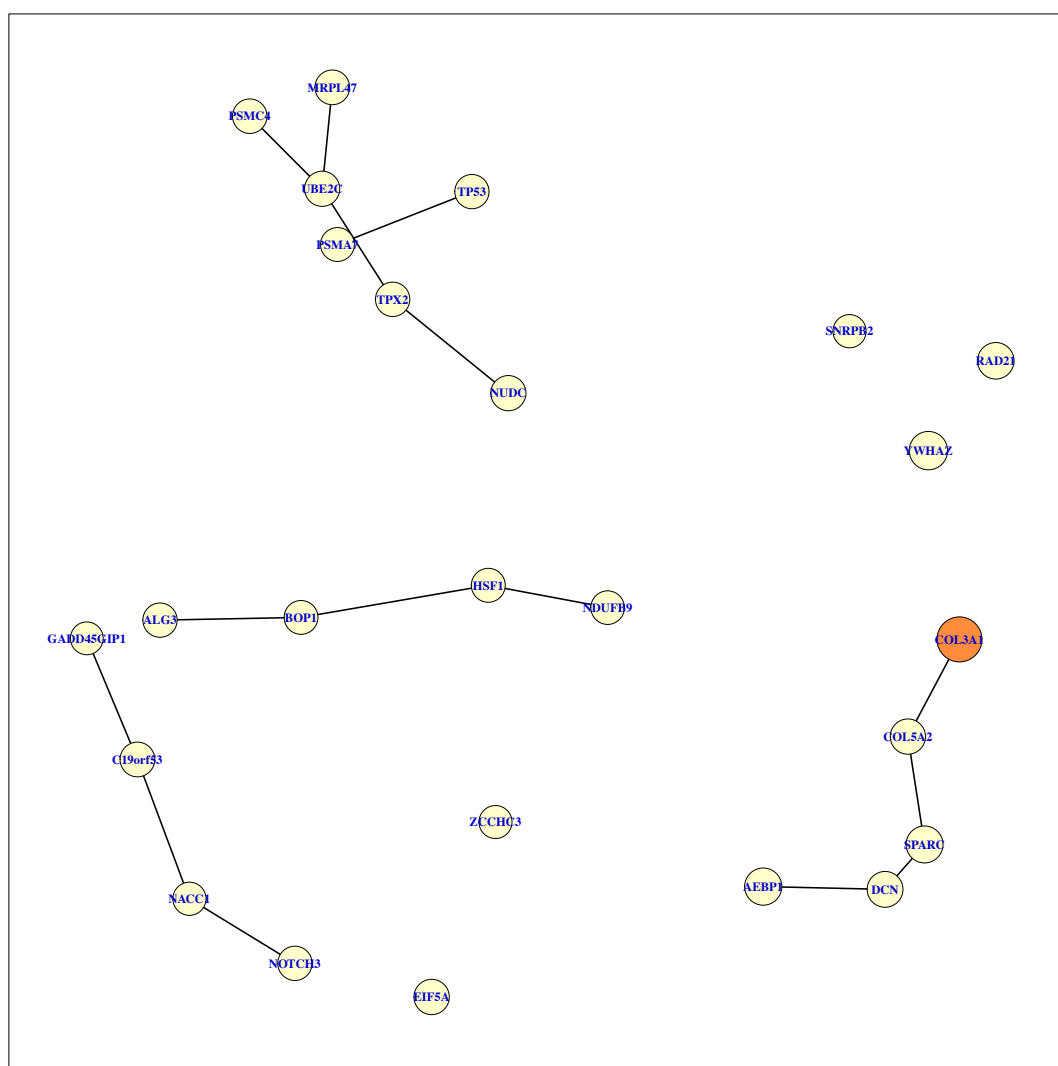


図 17 $k=13$ における遺伝子ネットワーク

$k = 13$ の基底では NUDC, TPX2, UBE2C などから成るネットワークと, NDUFB9, HSF1, BOP1, ALG3 から成るネットワーク, GADD45GIP1, C19orf53, NACC1, NOTCH3 から成

るネットワークが現れている (図 17). NUDC, TPX2, UBE2C などは細胞分裂に関わる遺伝子であり, その中でも UBE2C は, がんの進行に関与する遺伝子である. また, NDUFB9, HSF1, BOP1, ALG3, C19orf53, NOTCH3 は発がんなどに関わるがん遺伝子として同定されており, GADD45GIP1, NACC1 はアポトーシスに関与している. これらのことから, コラーゲン繊維と発がん・抑制に関わる遺伝子ネットワークであると解釈できる.

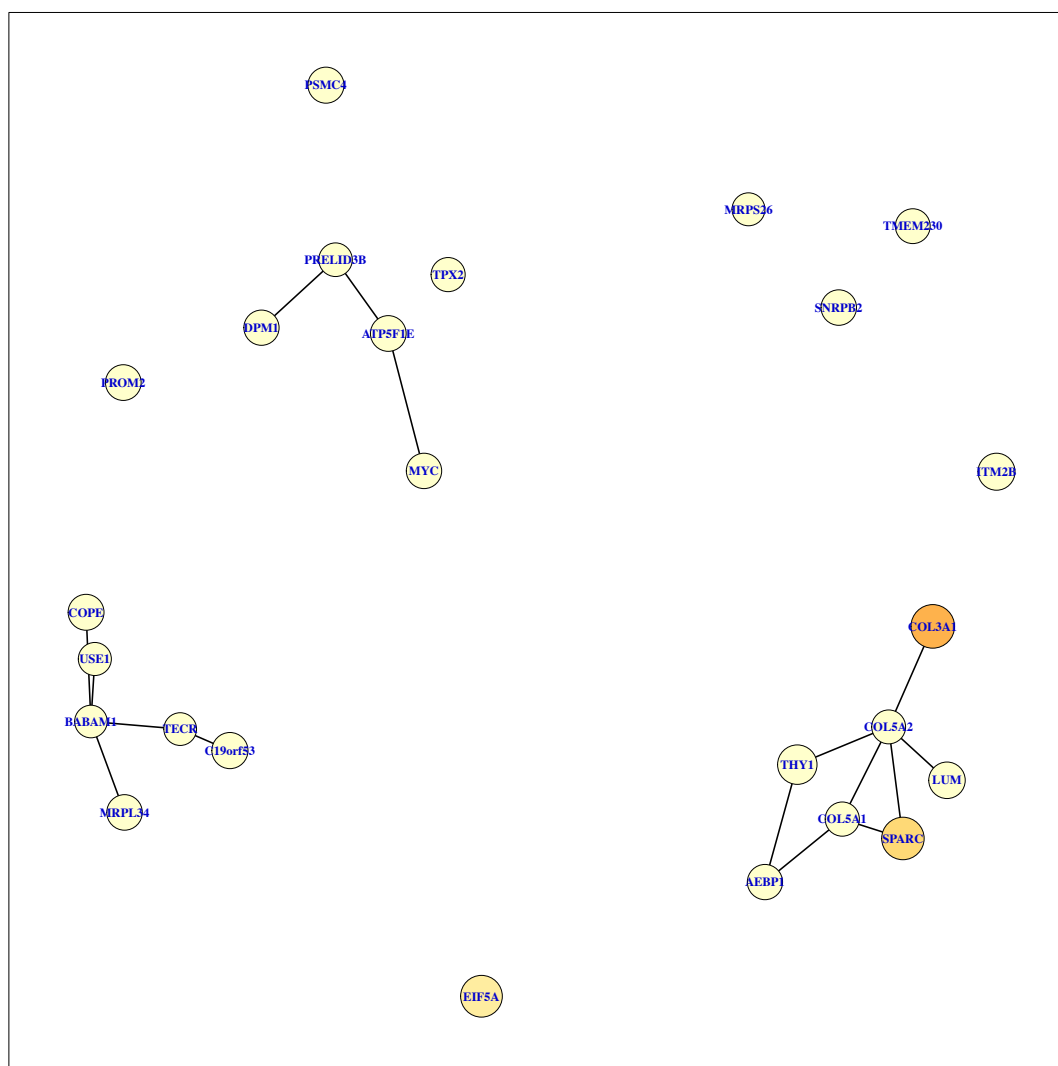


図 18 $k=29$ における遺伝子ネットワーク

$k = 29$ の基底に含まれるネットワークについて, MYC は数多くのがんに関わる発がん性のがん遺伝子であり, ATP5F1E や PRELID3B などはミトコンドリアの働きに関与する遺伝子出ある. がんはミトコンドリアの機能を利用するため, 生物学的知見に沿った関係性が現

れている。また、COPE や USE1 はゴルジ体の機能に関わる遺伝子であり、BABAM1 はユビキチンを認識する複合体の一部であり、コラーゲン繊維がユビキチンされた小胞体によってゴルジ体を經由し運搬する仕組みが知られている。これらのことから、コラーゲン繊維の輸送とがん細胞におけるエネルギー代謝に関わる遺伝子ネットワークであると解釈される。

以上の遺伝子ネットワークは、共通してコラーゲン繊維に関係する遺伝子が特徴的に現れており似た系であることが分かる。また、それぞれの基底では特徴的な遺伝子ネットワーク以外にも現れており、主な機能に関するネットワークとプラス α の機能に関わるネットワークが現れているのが見受けられる。このように、共通した遺伝子やその間でのネットワークを保持しつつ、細胞内における他機能に関する部分が現れるような結果が得られるのは従来法では得られないものである。

次にクラスタ 2 に属した基底群の遺伝子ネットワークを以下に示す。

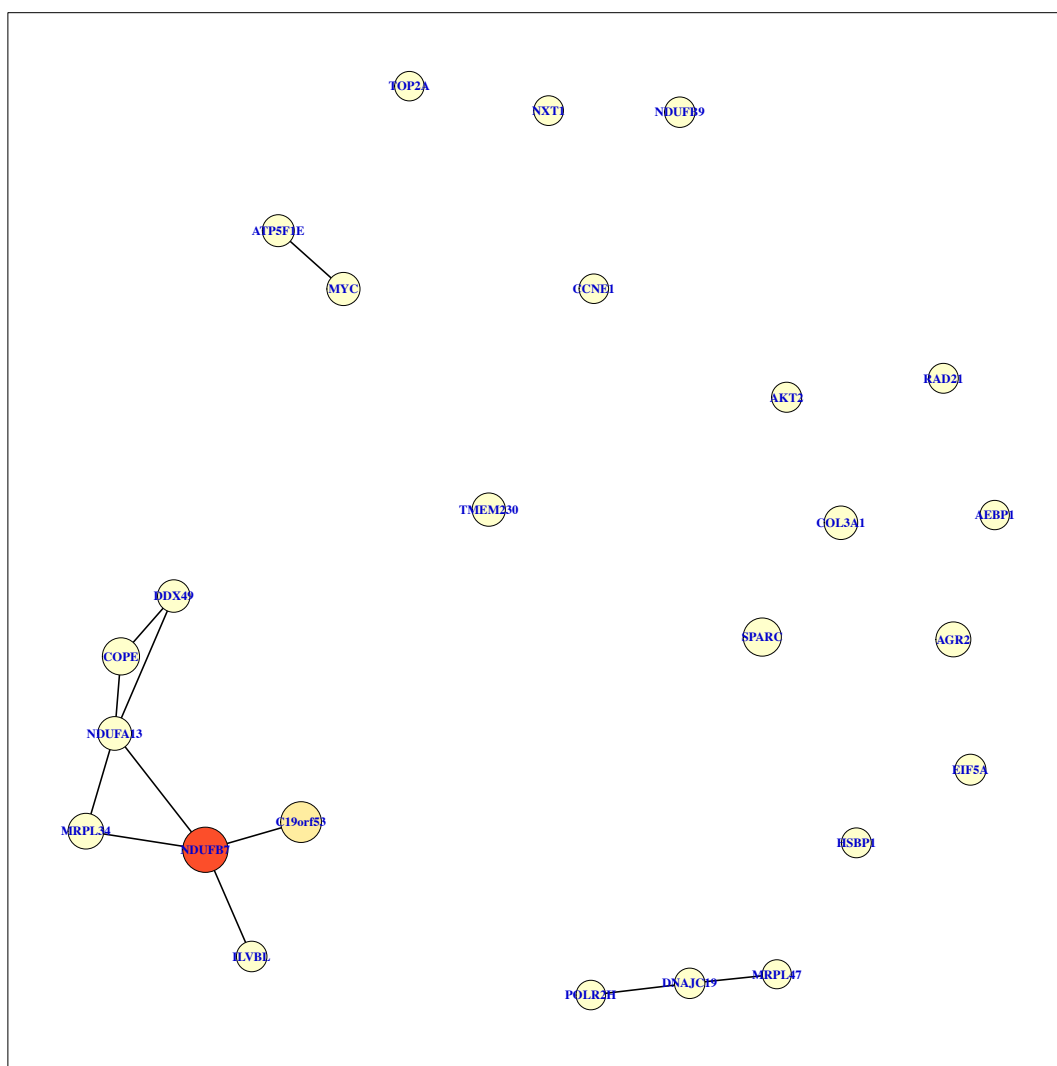


図 19 k=23 における遺伝子ネットワーク

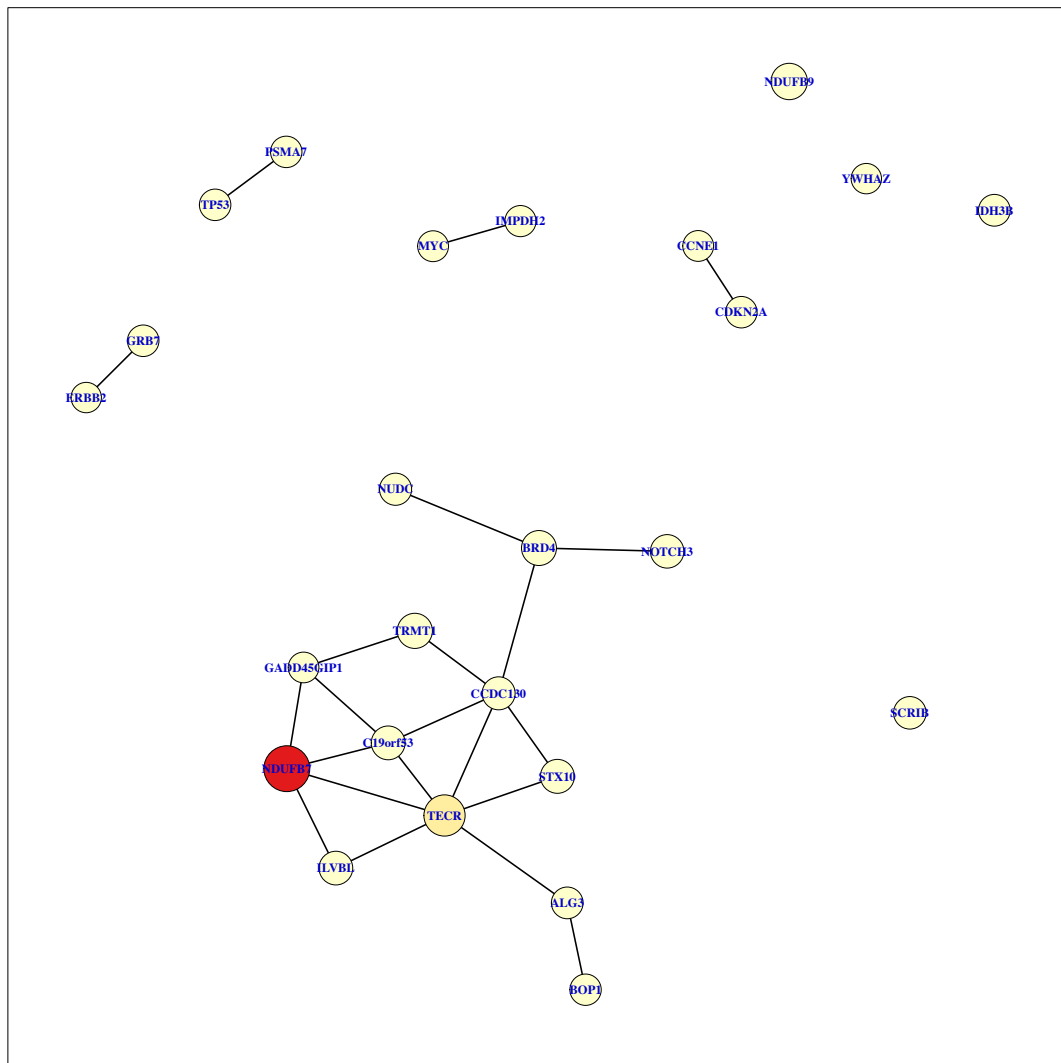


図 20 $k=33$ における遺伝子ネットワーク

これらの遺伝子ネットワークでは、NDUF7, C19orf53, ILVBL が共通部分として現れている。NDUF7 はミトコンドリア内膜に存在する複合体 I のサブユニットをコードしており、ILVBL は糖質代謝に関わるタンパク質をコードしている。C19orf53 は卵巣がんに関わる癌遺伝子として同定されており、がん細胞はミトコンドリアの機能に異常を来すことが知られている。このことから、これらの遺伝子ネットワークはがん細胞でのエネルギー代謝に関わるものと解釈することができる。

また、 $k = 23$ では NDUF13 や MRPL34 といったミトコンドリア内で働きを持つ遺伝子が連なっており、特に NDUF13 は腫瘍抑制因子としても働くことが知られており、NDUF7 がコードする複合体 I もアポトーシスを誘発する働きがあるとも推測されている。

これらのことから、がん細胞でのエネルギー代謝とミトコンドリア内におけるがん抑制応答に関わる遺伝子ネットワークであると解釈できる。

一方で、 $k = 33$ では TECR も特徴的に現れている。TECR は脂肪酸アシル CoA の代謝に関わるタンパク質をコードしており、ミトコンドリアと深い関わりがある。がんは解糖系を亢進させると同時に脂肪酸合成も盛んに行われていることが知られている。このことから、がん細胞でのエネルギー代謝と二次的な合成反応に関わる遺伝子ネットワークであると解釈できる。

次にクラスタ 3 に属した基底群の遺伝子ネットワークを以下に示す。

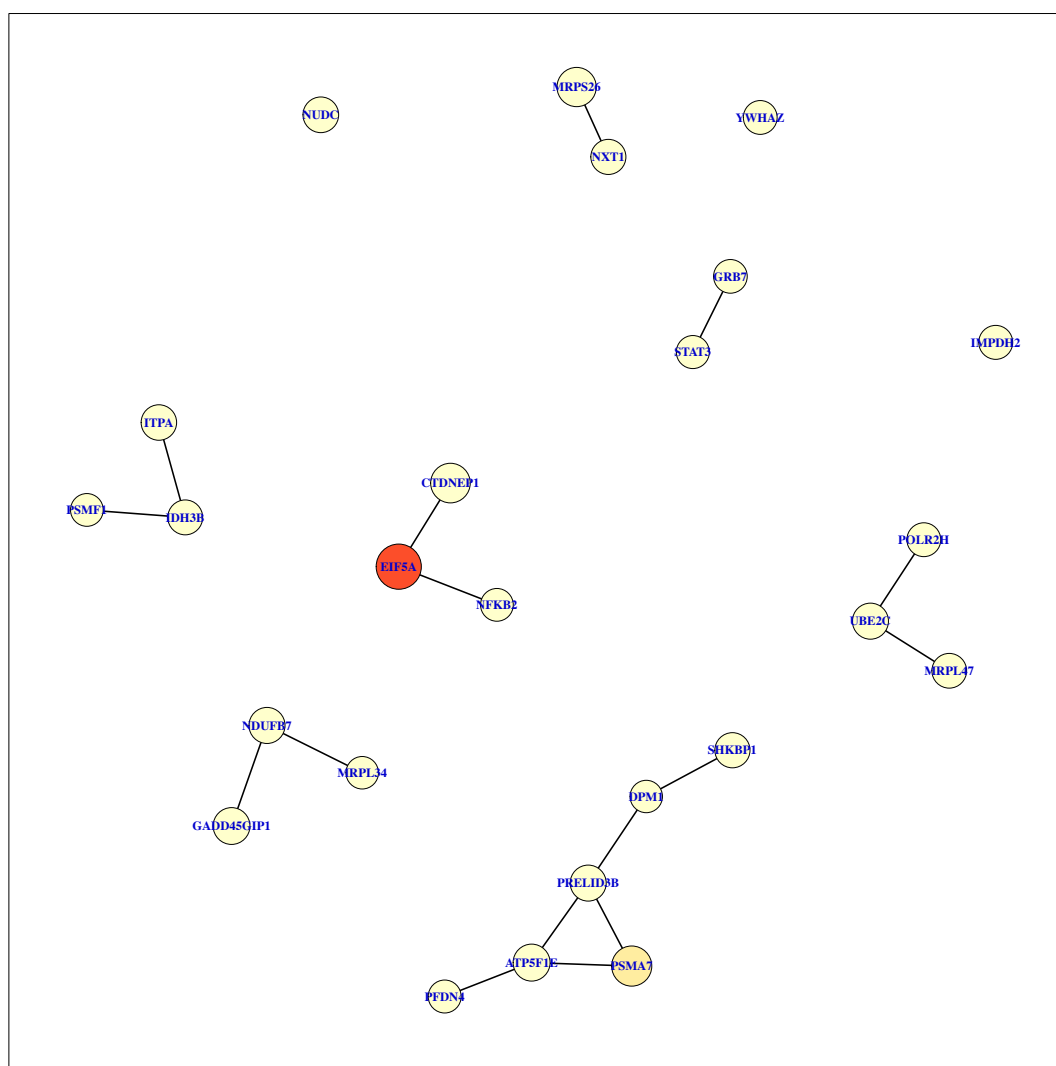


図 21 $k=19$ における遺伝子ネットワーク

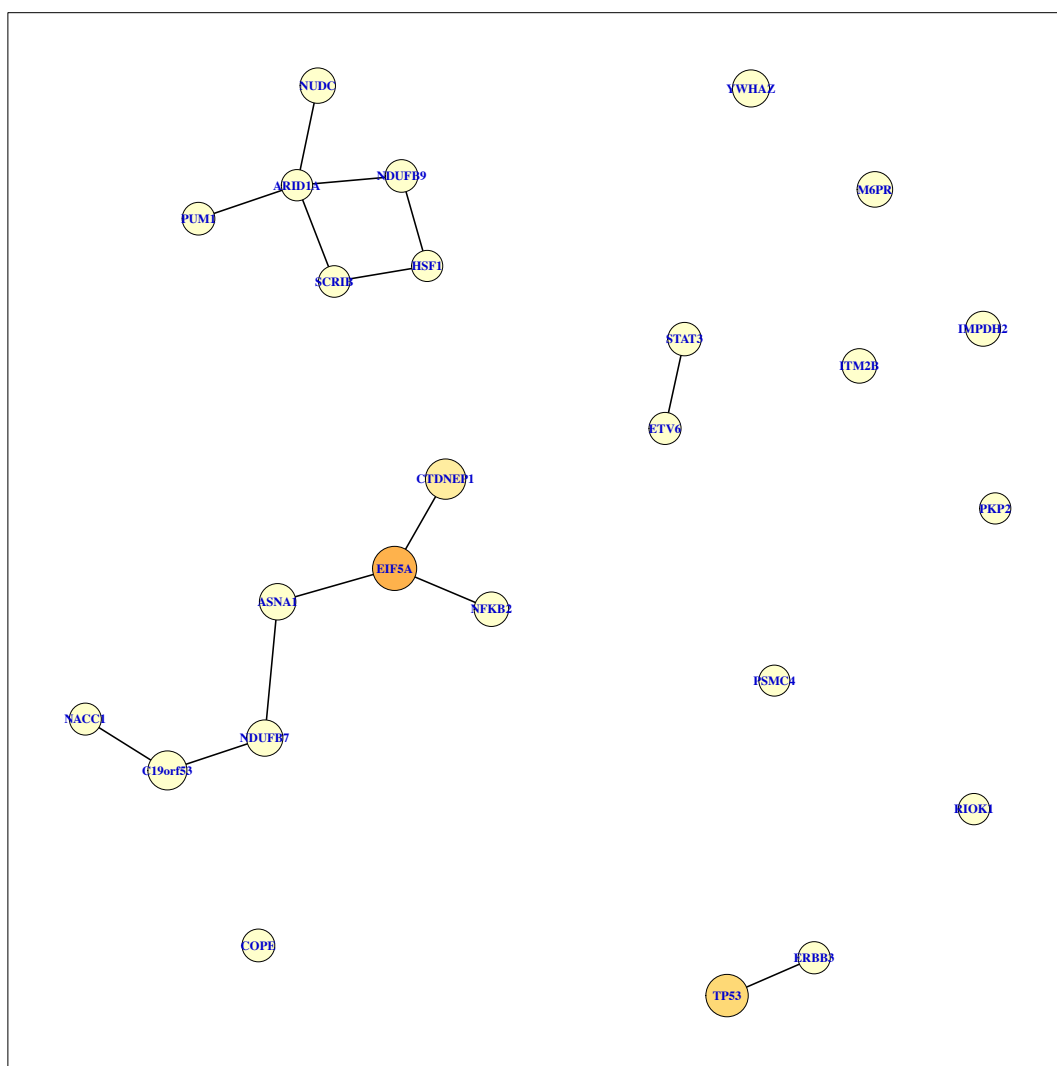


図 22 $k=28$ における遺伝子ネットワーク

$k = 19$ と 28 で得られた遺伝子ネットワークは EIF5A, CTDNEP1, NFKB2 を共通部分とし、特徴的な遺伝子として現れている。EIF5A は、細胞周期やストレス応答に関わる遺伝子であると同時に、がん細胞増殖を増強することが知られている。CTDNEP1 も細胞周期や有糸分裂に関わるタンパク質をコードしていることや、NFKB2 はがん細胞の不死身化、増殖・転移の増強をもたらすことから、がん細胞での増殖に関わる遺伝子ネットワークであると解釈される。

また、 $k = 19$ では、PSMA7 を含む遺伝子ネットワークも特徴的に現れている。この PSMA7 は結腸がんの肝転移に関わる遺伝子であることが知られている。一見すると卵巣がんに関わりがないように見られるが、卵巣がんはステージ II で子宮・卵管・直腸・膀胱に転

移することが知られている。これらのことから、PSMA7を含む遺伝子ネットワークはステージII以降で起こりうる転移で発現する遺伝子ネットワークではないかと推測され、がんの増殖と転移に関わる遺伝子ネットワークであると解釈できる。

次に、 $k = 28$ では、EIF5Aに続いてASNA1, NDUF7, C19orf53, NACC1が連なっている。ASNA1は小胞体に関係するタンパク質をコードしており、がん増殖とミトコンドリアに深い関わりがある。がん増殖では小胞体による周囲の正常細胞をがん化させる仕組みが知られており、ミトコンドリアと小胞体が接触することで細胞内の老廃物や病原体を分解するオートファゴソームを形成することが知られている。また、C19orf53は卵巣がんのがん遺伝子として同定されており、NACC1は腫瘍抑制因子を抑制し、がんの進行を助けることが知られている。他にも、TP53はがん抑制因子として有名であり、ERBB3は様々ながんでの異常発言が報告されている。これらのことから、 $k = 28$ では、がんの増殖と抑制に関わる遺伝子ネットワークであると解釈できる。

以上のように、基底のクラスタリングから、クラスタ1, 2, 3に属する基底はそれぞれがある機能を基盤とする遺伝子ネットワークであることが示せた。その一方で、クラスタの中に含まれる各基底は、それぞれがプラス α の機能に関係している遺伝子ネットワークを含むことが分かった。しかし、クラスタ4は同一の機能を有している訳ではないため、それぞれが独立した基底であると考えられる。そこで、以下よりクラスタ4に属する基底についての解釈を一つずつ述べる。

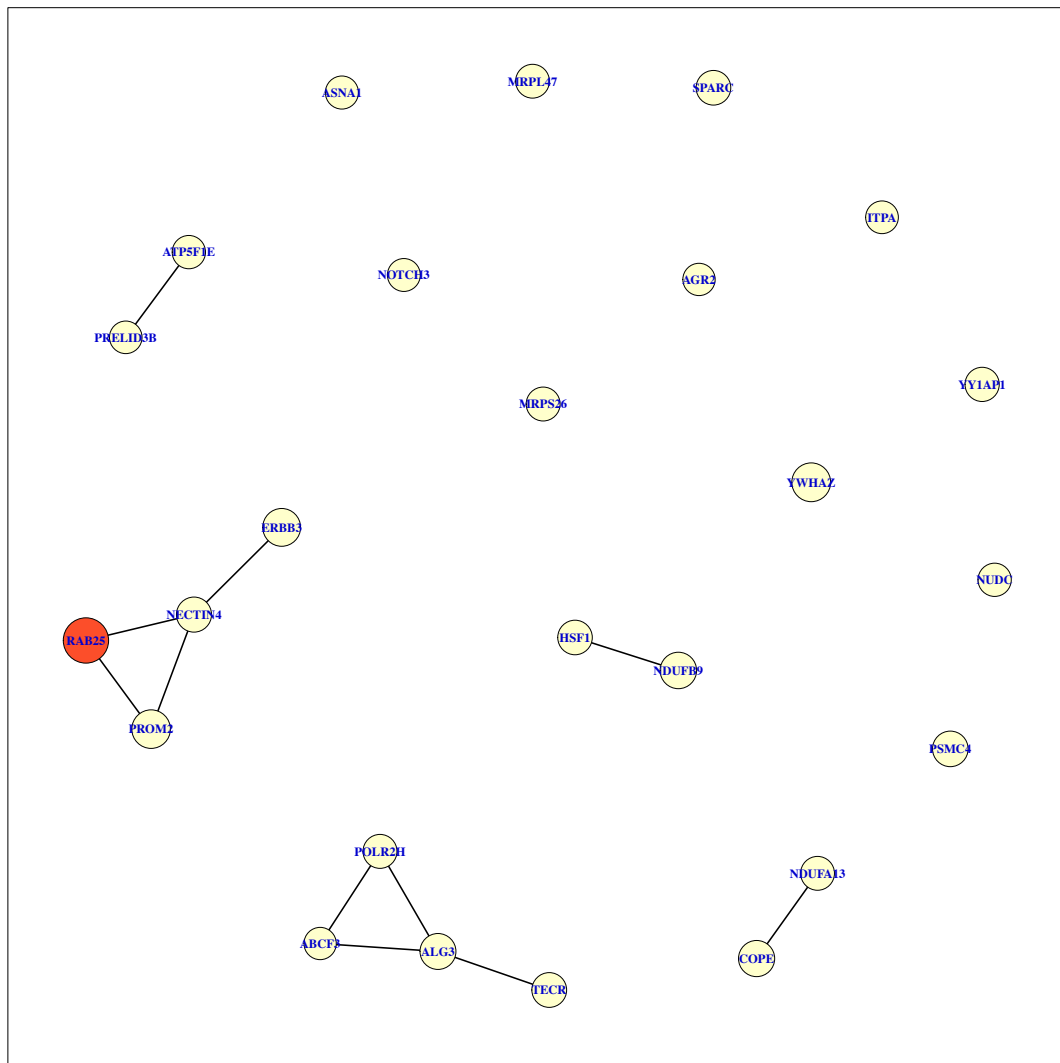


図 23 $k=20$ における遺伝子ネットワーク

$k = 20$ では RAB25, NECTIN4, ERBB3, PROM2 のネットワークが特徴的である。これらの遺伝子は細胞の上皮での機能に関わるたんぱく質をコードしている(図 23)。また, TECR, ABCF3 も膜輸送に関わる遺伝子である。RAB25, ERBB3, ALG3 はがん遺伝子としても知られるが、特に ALG3 は扁平上皮がんに関わる遺伝子である。これらのことからがん細胞と細胞上皮における働きに関係した遺伝子ネットワークであると解釈できる。

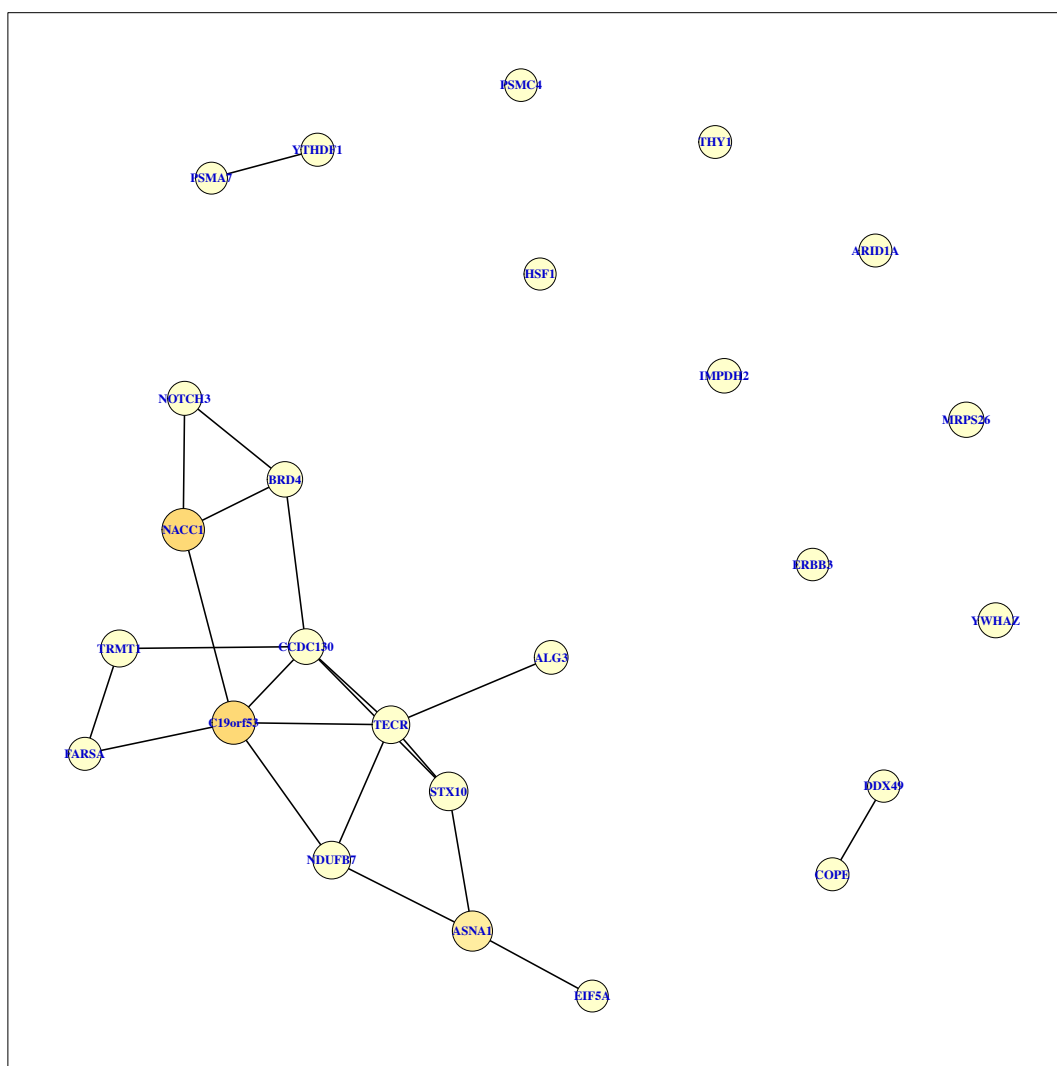


図 24 $k=12$ における遺伝子ネットワーク

$k = 12$ では C19orf53, ASNA1, NACC1 などが特徴的に現れた遺伝子ネットワークがある (図 24). IMPDH2, C19orf53, NOTCH3, BRD4 はがん遺伝子として知られ, 特に C19orf53 と BRD4 は卵巣がんに関わる癌遺伝子として同定されている. また, ALG3, TECR, ASNA1 はタンパク質輸送に関わるタンパク質をコードしており, NACC1, NDUFB7 にはアポトーシスに関わる機能を有することから細胞の小胞体ストレス応答に関わる遺伝子ネットワークであると解釈できる.

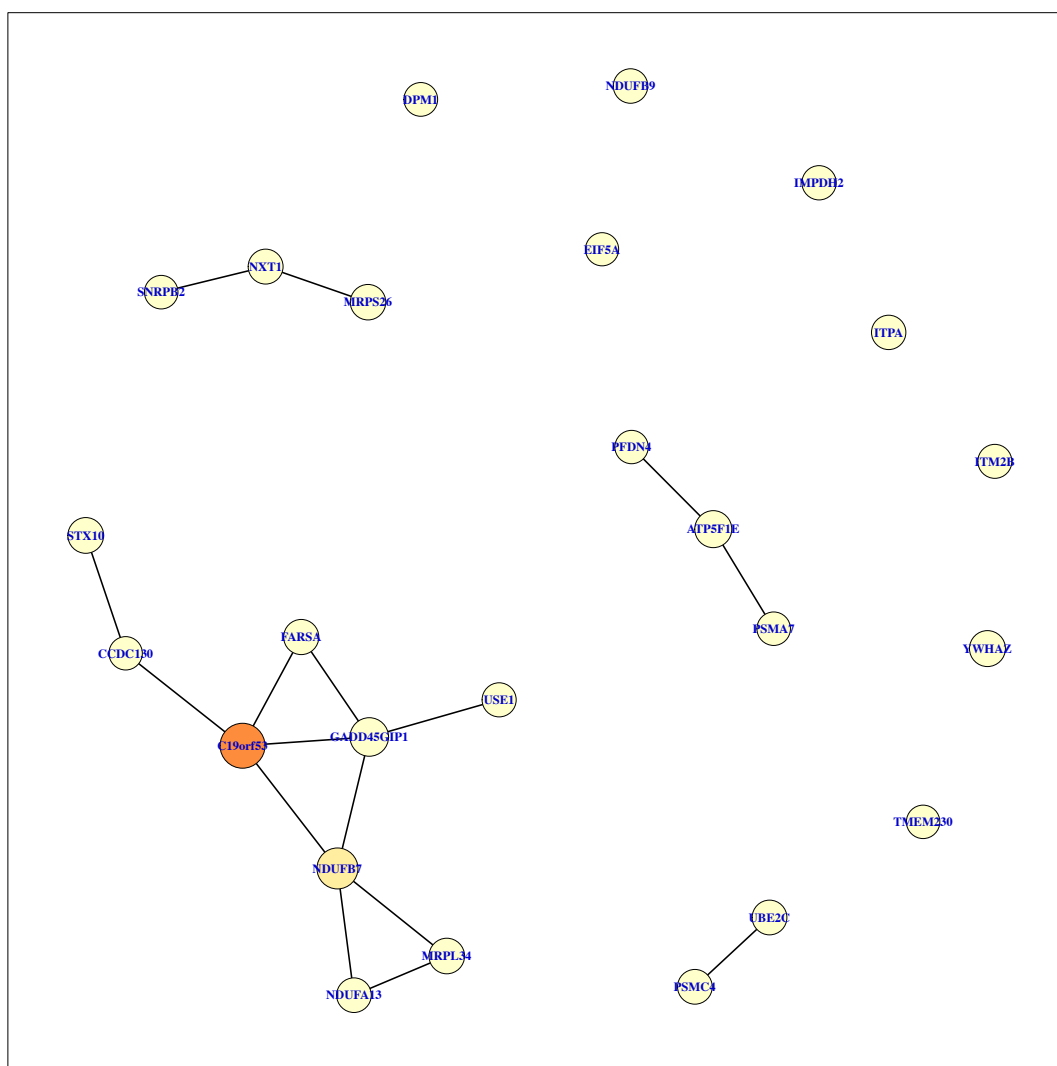


図 25 $k=35$ における遺伝子ネットワーク

$k = 35$ では C19orf53, NDUF7 などを含む遺伝子ネットワークが特徴的に現れている (図 25). C19orf53 は卵巣がんに関わる癌遺伝子として同定されており, FARS4, GADD45GIP1 は細胞周期に関わる遺伝子であり, NDUF7, NDUF13, MRPL34 はミトコンドリアの機能に関わる遺伝子である. これらのことから, がん細胞における細胞周期とミトコンドリアに関わる遺伝子ネットワークであると解釈される.

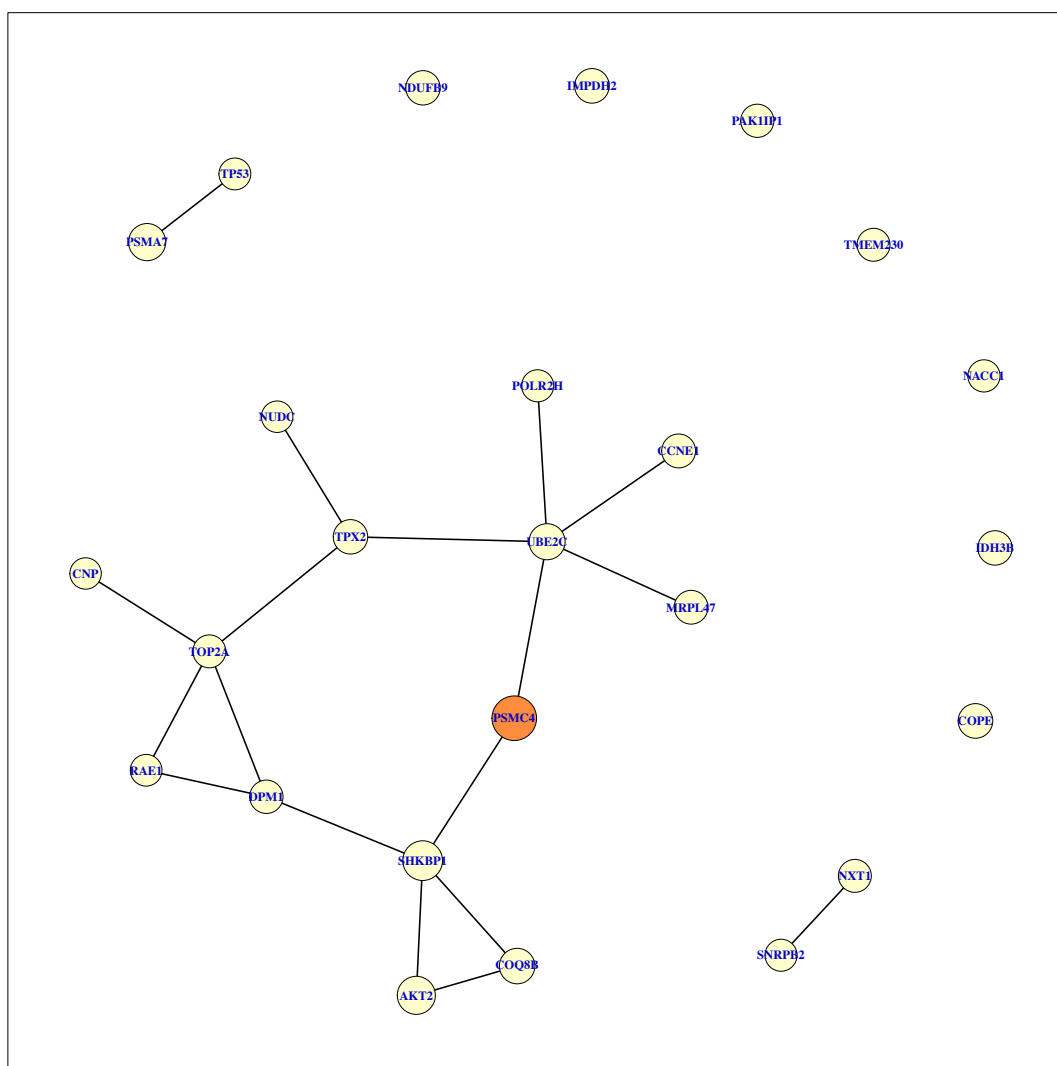


図 26 $k=1$ における遺伝子ネットワーク

$k = 1$ では PSMC4, UBE2C, TPX2, SHKBP1 などを含むネットワークが特徴的である。TPX2 や UBE2C は遺伝子発現や細胞周期に関わる遺伝子である (図 26)。また、UBE2C は分解するたんぱく質に標的化をするユビキチン化を施す機能を持ち、プロテアソームの機能に関わる PSMC4 と関係がある。SHKBP1 や AKT2 は子宮頸がん・卵巣がんを発生させる癌遺伝子として同定され、子宮頸がんではプロテアソームを利用した癌抑制たんぱく質の分解によるアポトーシス阻害が示唆されている。このことから、がん細胞における細胞分裂と癌抑制反応の阻害に関わる遺伝子ネットワークであると解釈される。

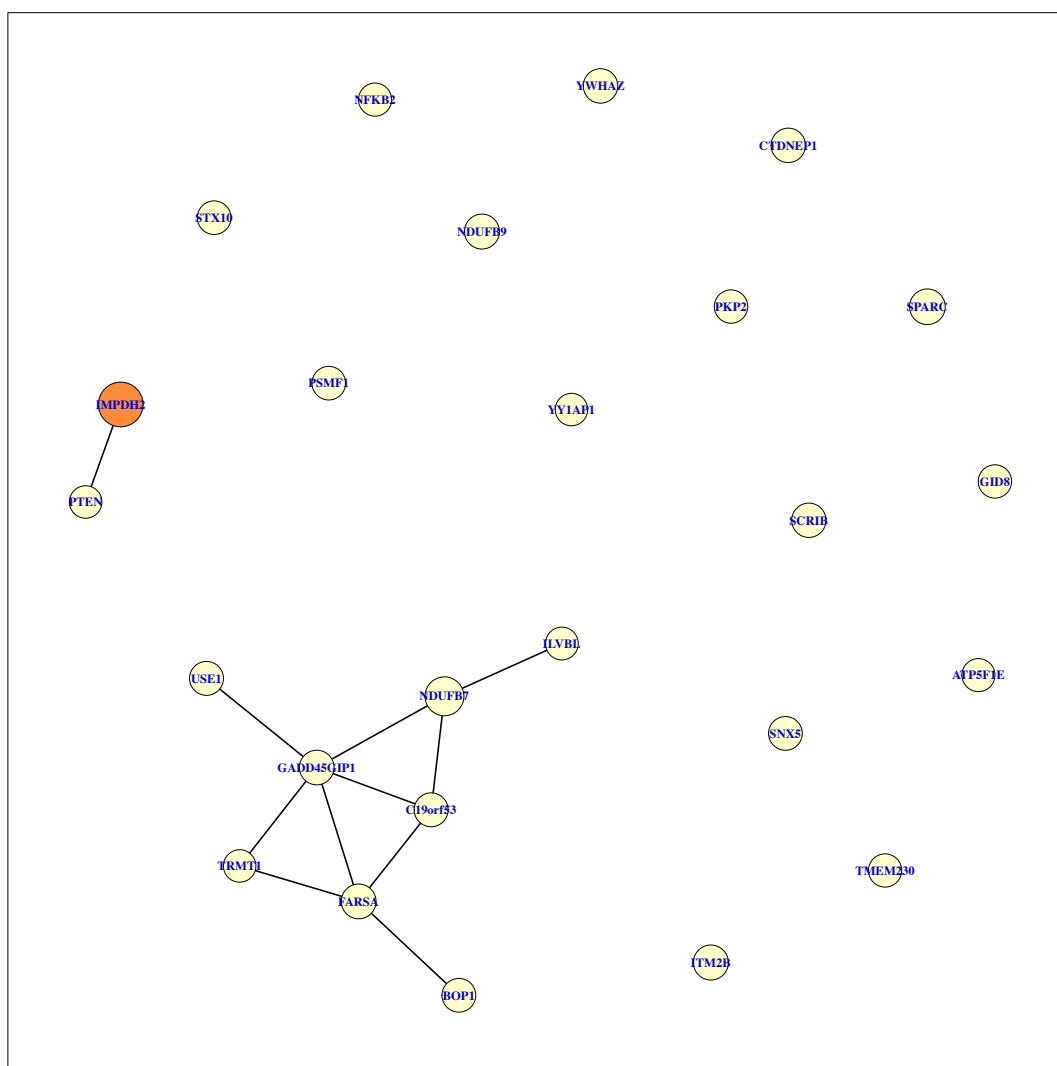


図 27 $k=8$ における遺伝子ネットワーク

$k = 8$ では細胞周期の調節や DNA・RNA の合成に関わる PTEM や FARS4, IMPDH2 が特徴的に現れており，それに伴う tRNA の機能に関わる TRMT1 も現れていた．他にはミトコンドリアでのエネルギー代謝やがん遺伝子，がん抑制が現れていた．これらのことから発がんや抑制，細胞分裂に関わる遺伝子ネットワークであると解釈できる．

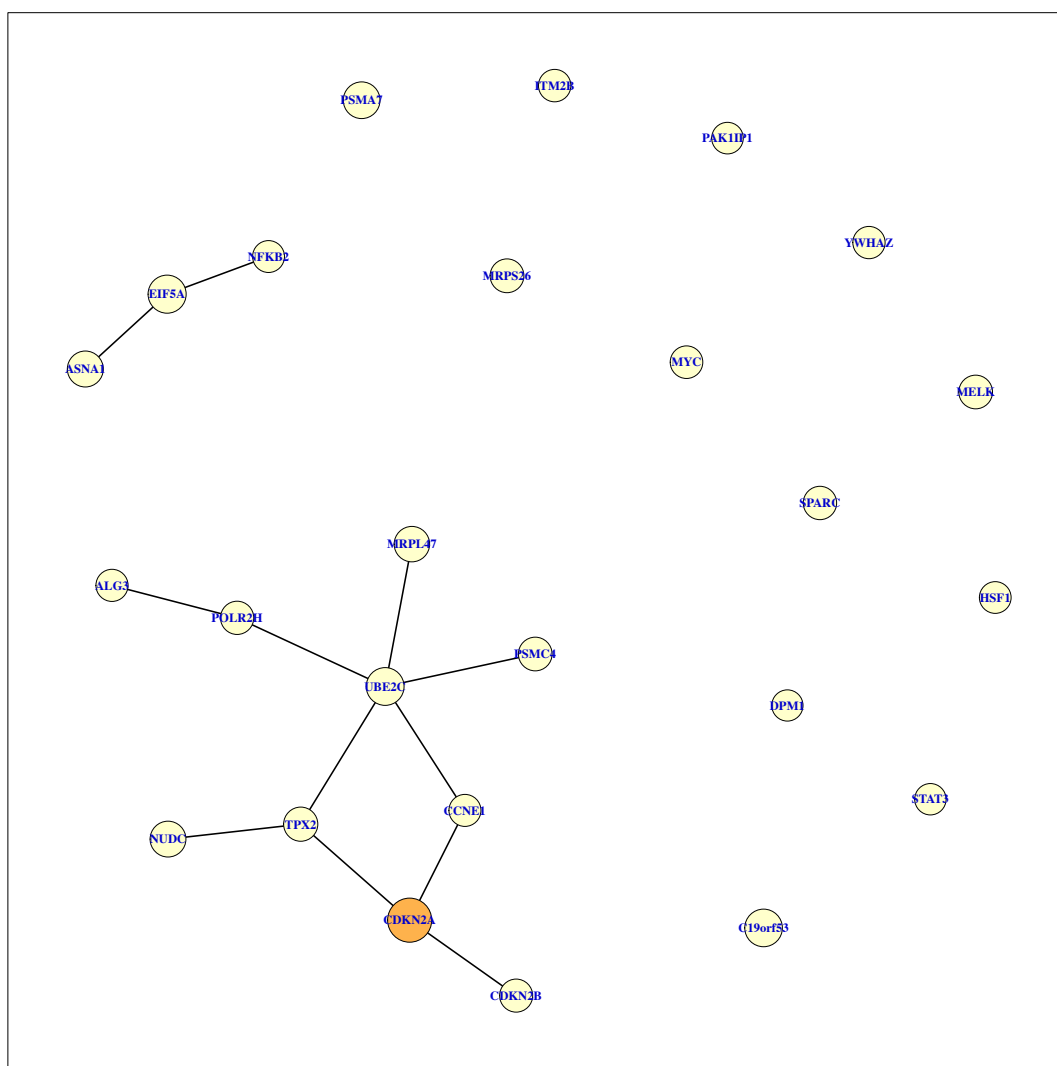


図 28 $k=36$ における遺伝子ネットワーク

$k = 36$ では、CDKN2A, CDKN2B, CCNE1, TPX2, UBE2C などを含む遺伝子ネットワークが主に現れている。これらは細胞周期や細胞分裂に関わる遺伝子であるため、そのネットワークであると解釈できる。また、CDKN2A は細胞周期を調節することで腫瘍抑制因子として作用することが知られており、がん細胞では TP53 について発現が抑制されていることが確認されている。CCNE1 は CDK キナーゼと作用し合うことが知られている。TPX2 や UBE2C はがんに関わる遺伝子としても知られている。

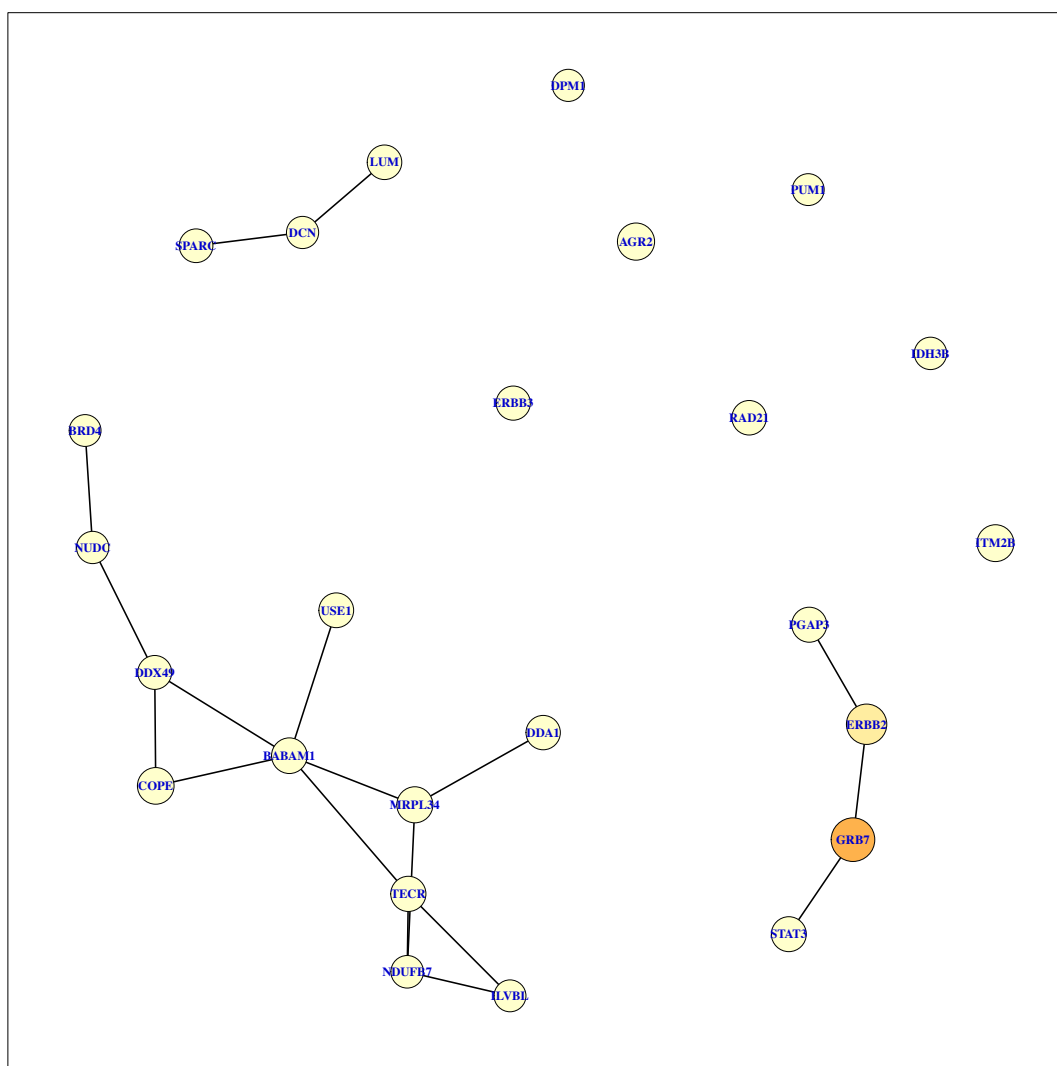


図 29 $k=31$ における遺伝子ネットワーク

$k = 31$ では GRB7, ERBB2, STAT3, PGAP3 が主なネットワークとして現れている。GRB7, ERBB2 は両方ともシグナル伝達に関わり、細胞の様々なプロセスに関与している。STAT 3 はシグナルに応答して増殖やアポトーシスなどの細胞プロセスに関わり、PGAP3 もグリコシルホスファチジルイノシトール(glycosylphosphatidylinositol, GPI)に特異的な酵素をコードしており、GPI タンパクは受容体の働きに関与する系も含まれている。これらのことからシグナル伝達に関わる機能を有する遺伝子ネットワークであると解釈できる。

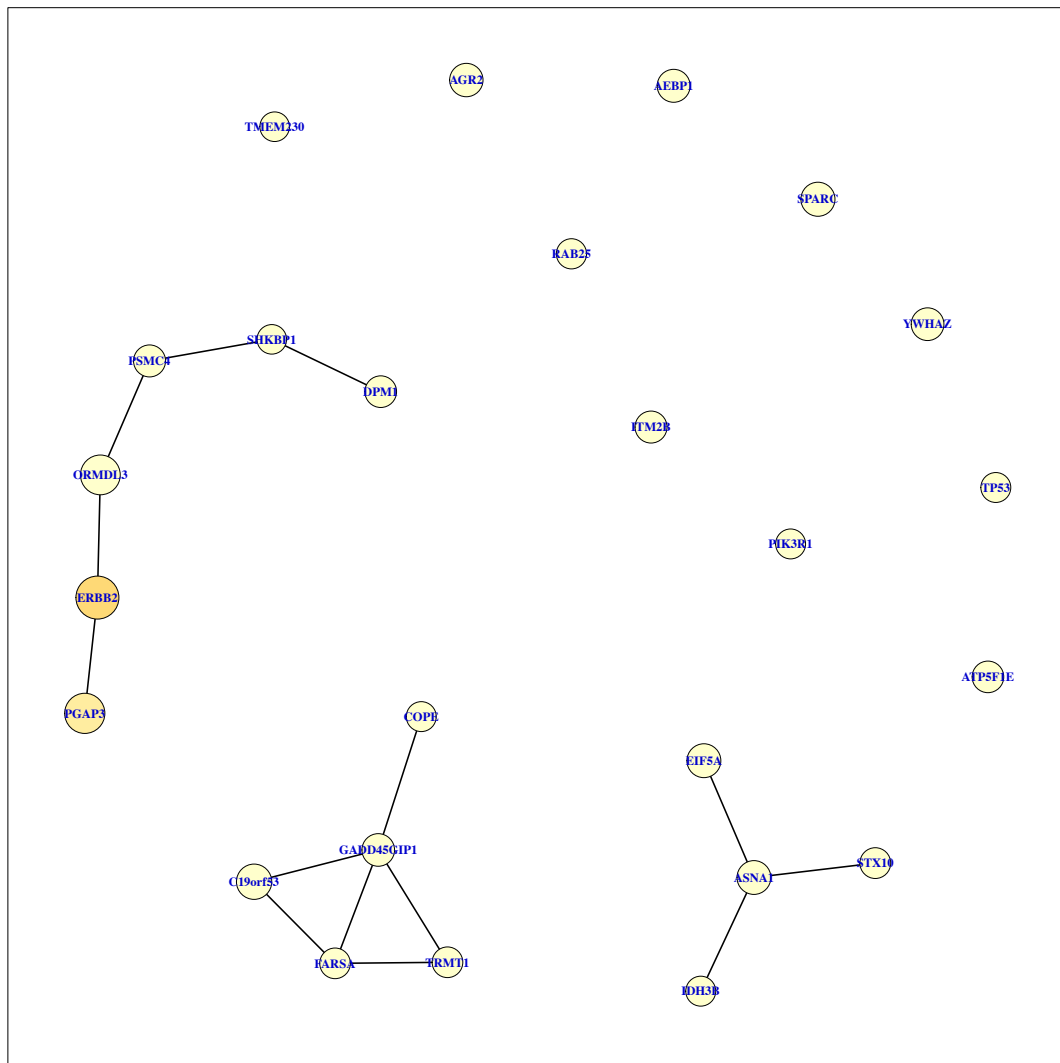


図 30 k=39 における遺伝子ネットワーク

k=39 では ERBB2, PGAP3, ORMDL3, PSMC4, SHKBP1, DPM1 が主なネットワークとして現れている。ERBB2, PSMC4 はシグナル伝達に関わり、SHKBP1 はイオンチャンネルに関与している。いずれも細胞内伝達の機構に関わるものである。特に ERBB2 の過剰発現は、乳房および卵巣の腫瘍を含む多数の癌において報告されており、PSMC4 がコードするタンパク質は、細胞の癌抑制蛋白である p53 や pRB などを分解し、アポトーシス阻害や細胞周期攪乱をきたすことが知られている。この異常が蓄積されることで、子宮頸部の細胞は異型性から癌化に至る重要なプロセスであると考えられている。これらのことからがん発生とシグナル伝達に関わる遺伝子ネットワークであると解釈できる。

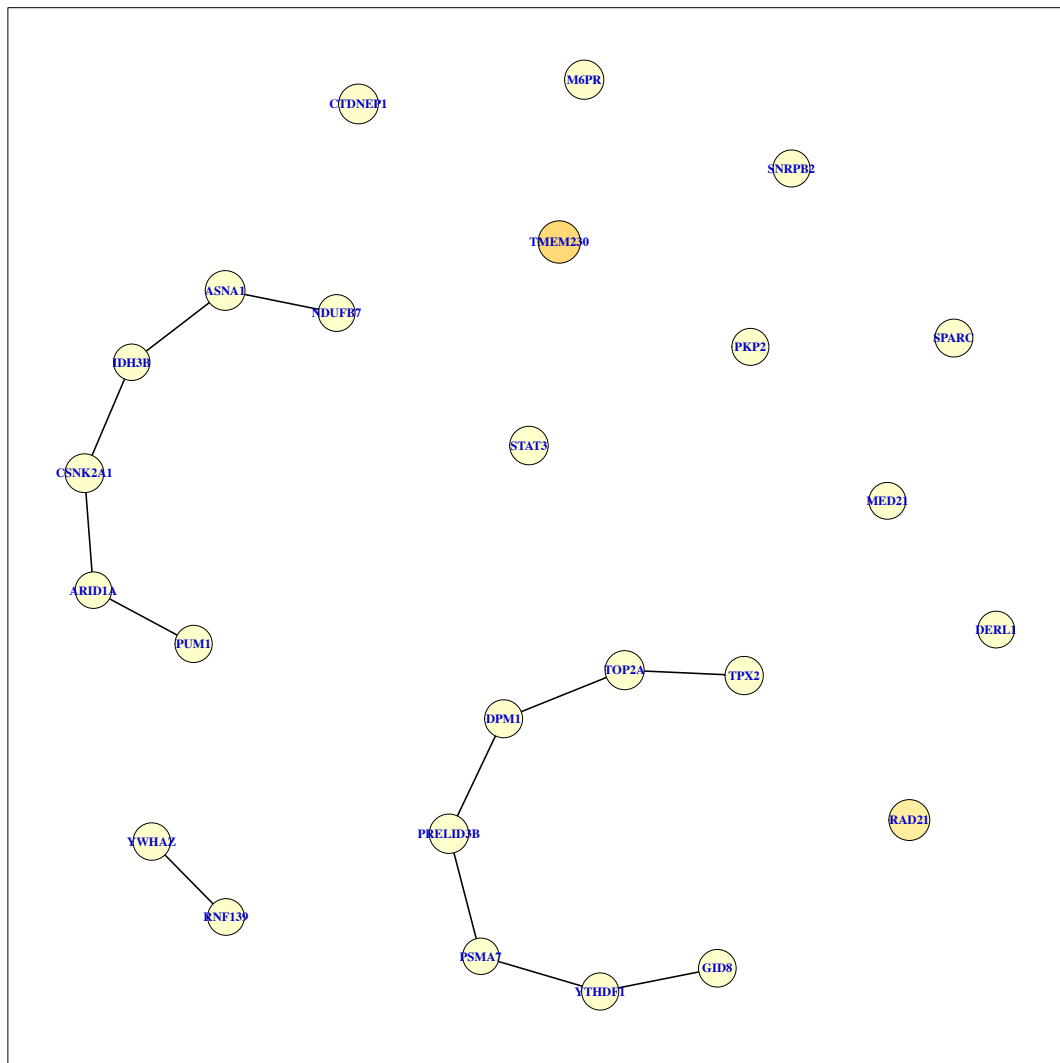


図 31 $k=11$ における遺伝子ネットワーク

$k = 11$ では、TMEM230 はシナプス小胞に関わる遺伝子であり、TMEM230 の疾患がパーキンソン病に関わることが明らかになっている。加えて、子宮頸がんの治療によってパーキンソン病のリスクが上がることも報告されている。RAD21 はコヒーシンに関わる遺伝子であり、RAD21 の異常によってシナプスの形成に異常を来すことも報告されている。このことから、特徴的に現れている2つの遺伝子は脳神経に関わりがあることが確認されたが、パーキンソン病自体未解明な部分が多く、二つの関係性を示す文献は見当たらなかった。一方で、ASNA1 はタンパク質輸送に関わるタンパク質をコードしており、IDH3B は細胞が活動するときに必要なエネルギーを生み出す代謝を触媒する働きがある。CSNK2A1 は細胞周期や有糸分裂などといった働きのほか、発がん性タンパク質をリン酸化することでも知

られた癌遺伝子である．タンパク質輸送やリン酸化にもエネルギーが必要とされることから，それぞれの遺伝子の相互作用が解釈できる．また，ARID1A は卵巣がんや子宮内膜がんを抑制する複合体タンパク質の形成を促進する抑制遺伝子として知られ，PUM1 は翻訳調節因子をコードしている遺伝子である．これらのことから，それぞれの遺伝子が関係性のある機能を有し，相互作用し合っていることが解釈できる．

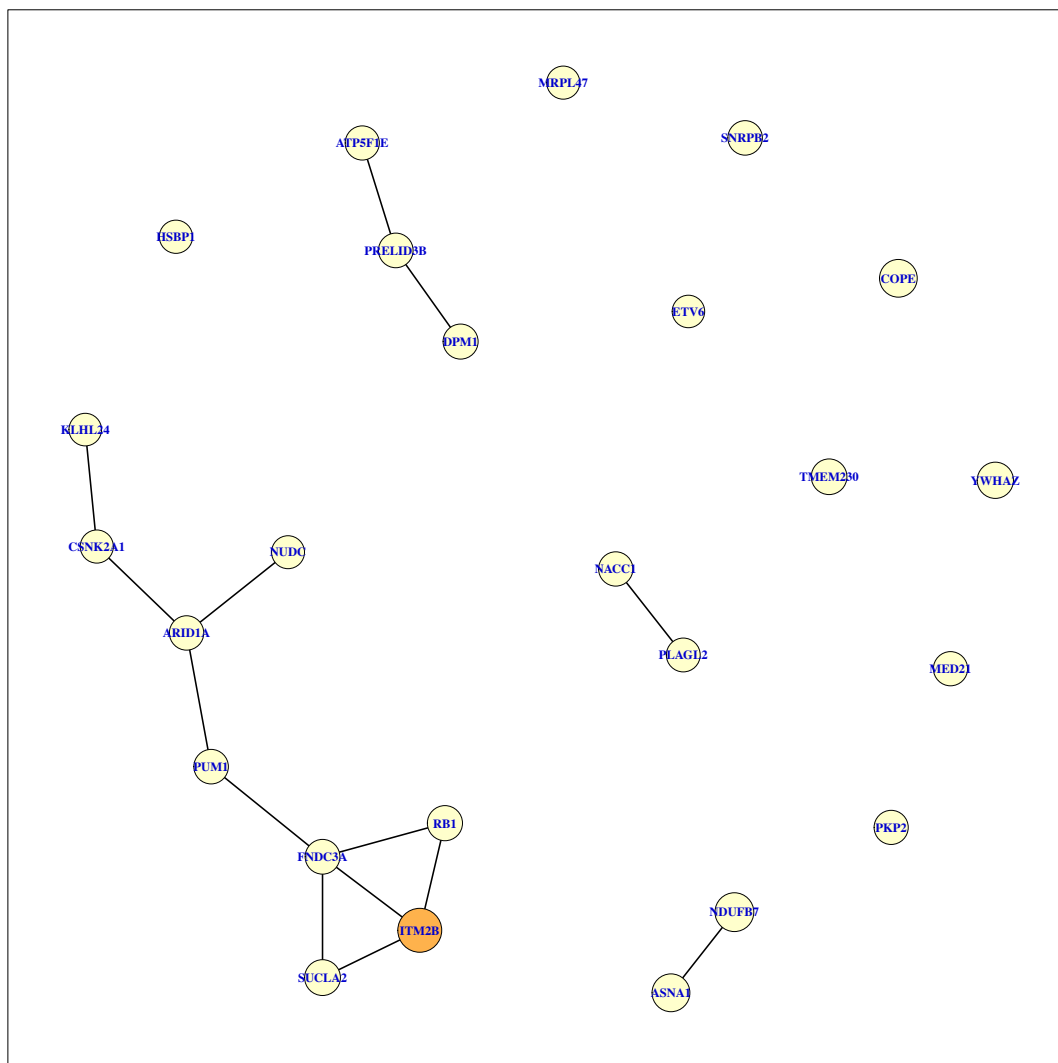


図 32 $k=38$ における遺伝子ネットワーク

$k = 38$ では ITM2B, SUCLA2, RB1, FNDC3A などを含む遺伝子ネットワークが主に現れている．ITM2B は脳神経系に関わるたんぱく質をコードしており，SUCLA2 は心臓，脳，骨格筋などの異化組織で主に発現していることが知られている．また，脳内において

SUCLA2 は神経細胞でのみ発現することも確認されている．RB1 や FNDC3A の異常は脳腫瘍で確認されており，これらの遺伝子が脳神経に関わるものであると解釈することができる．

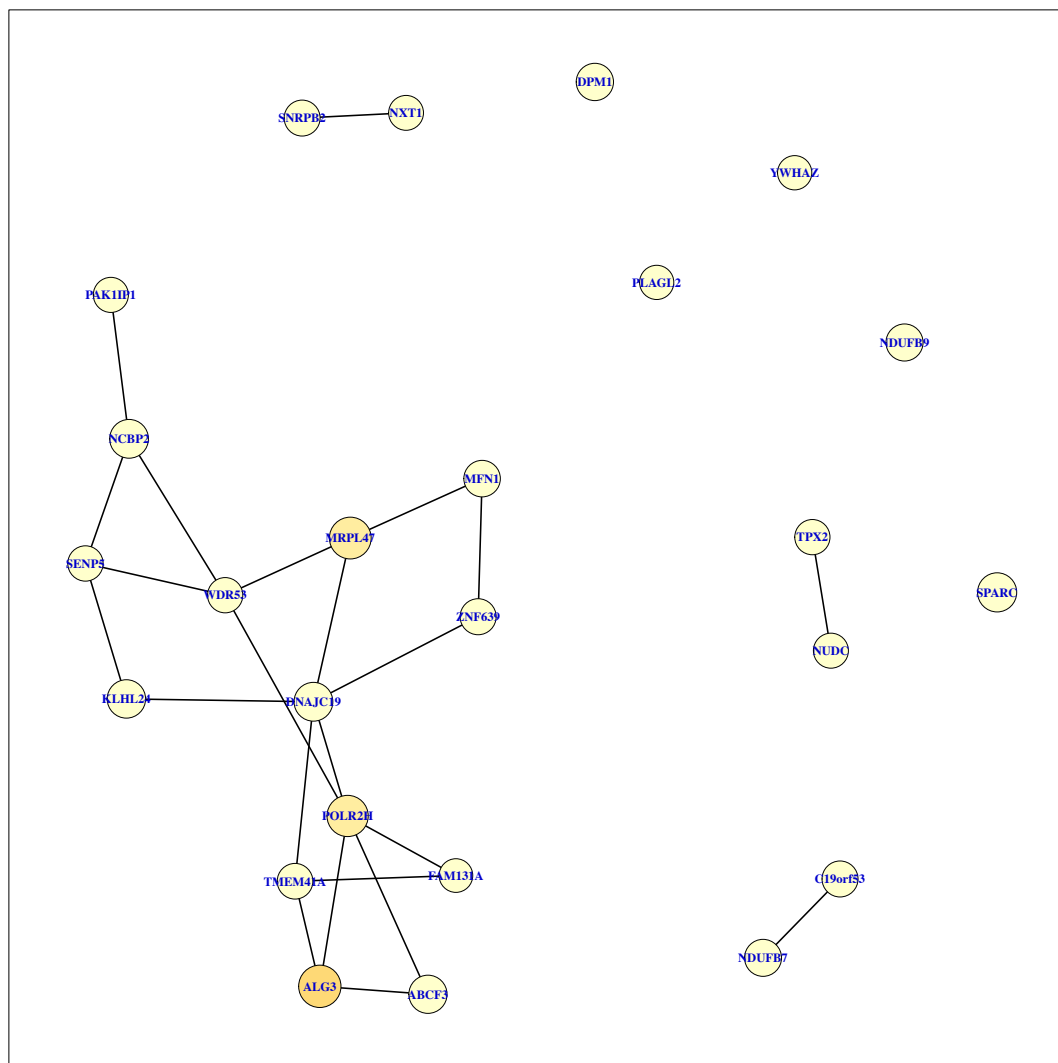


図 33 $k=34$ における遺伝子ネットワーク

$k = 34$ では ALG3, POLR2H, MRPL47, DNAJC19, ZNF639 などを含む遺伝子ネットワークが主に現れている．ALG3 はマンノース付加を触媒するタンパク質をコードし，乳がんや子宮頸がんに関連があることが知られている．また，マンノースはがんを抑制することが明らかになっている．POLR2H はタンパク質合成に必要なポリメラーゼ II をコードしており，MRPL47, DNAJC19 はミトコンドリアの機能に関わる．ZNF639 はがん細胞において，発

現の増進・過剰が確認されている。これらのことから、がん細胞の影響による細胞プロセスでの各遺伝子の増進に関する遺伝子ネットワークであると解釈できる。

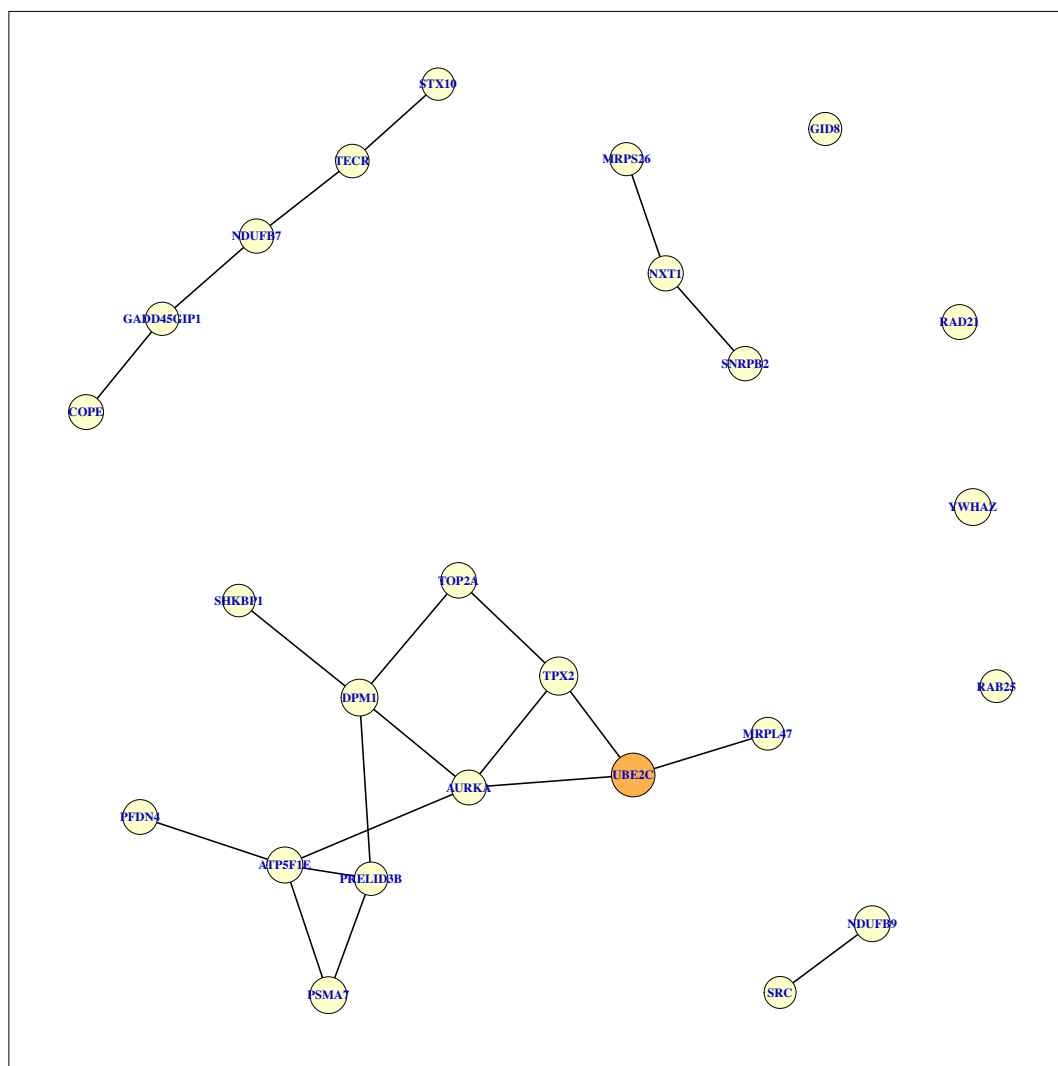


図 34 $k=6$ における遺伝子ネットワーク

$k = 6$ では細胞分裂・細胞周期に関わる遺伝子 UBE2C, TPX2, AURKA などが特徴的に現れており、ミトコンドリアの機能に関わる ATP5F1E や PELID3B など現れていた。これらから、細胞分裂とミトコンドリアに関わる遺伝子ネットワークであると考えられる。

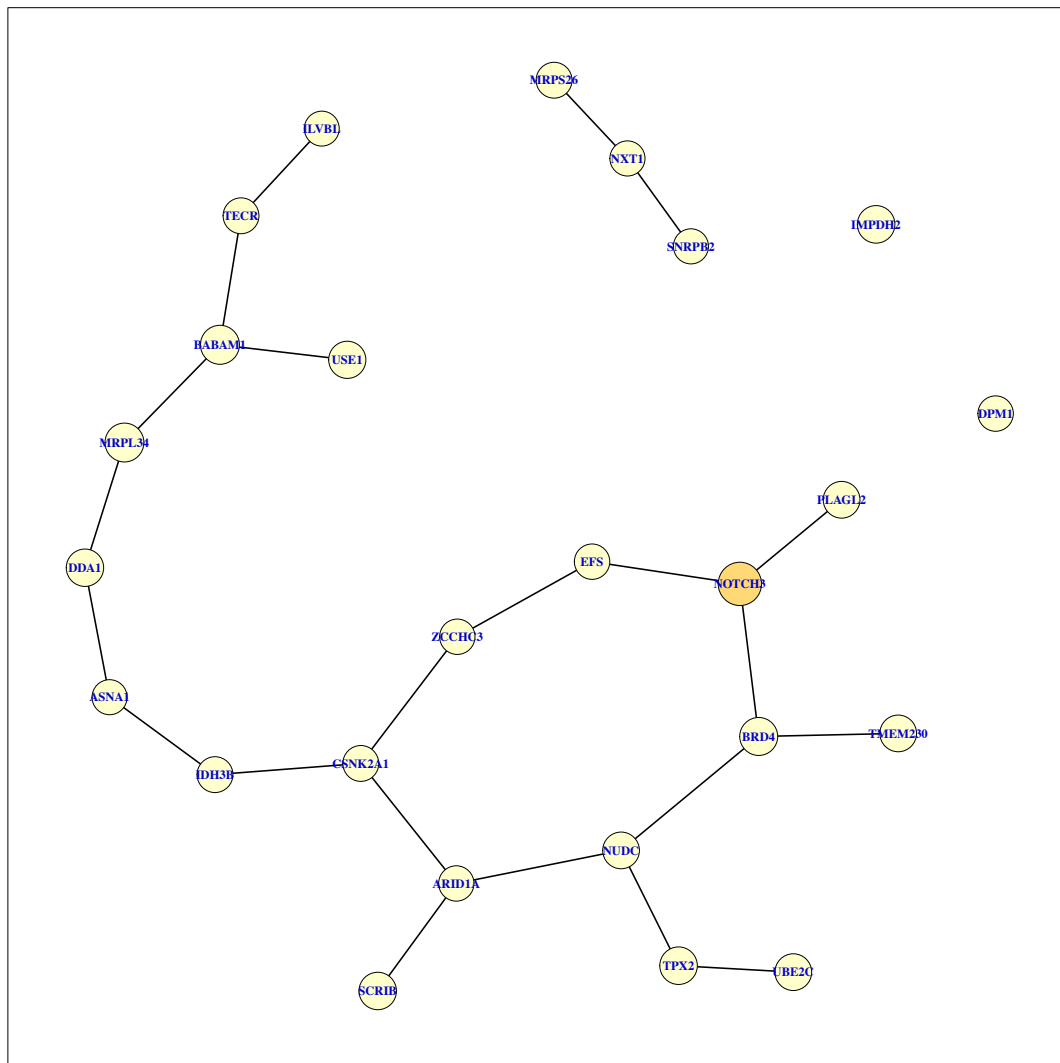


図 35 $k=27$ における遺伝子ネットワーク

$k = 27$ では NOTCH3, PLAGL2, BRD4, EFS, ZCCH3 などを含む遺伝子ネットワークが主に発現している. NOTCH3 は Notch ファミリータンパクをコードしており, 細胞間相互作用, 胚形成, および組織への関与を調節する遺伝子である. PLAGL2 は細胞の形質転換およびアポトーシスに関与し, BRD4 がコードするタンパク質は転写活性化補助因子である. EFS は細胞接着, 分化増殖, 癌化, 細胞内シグナル伝達において重要な役割を担い, ZCCH3 は核酸結合タンパク質はである tRNA をコードしており, 生体内での多様な役割を担っている. これらのことから様々な細胞プロセスに関わる遺伝子ネットワークであると解釈できる.

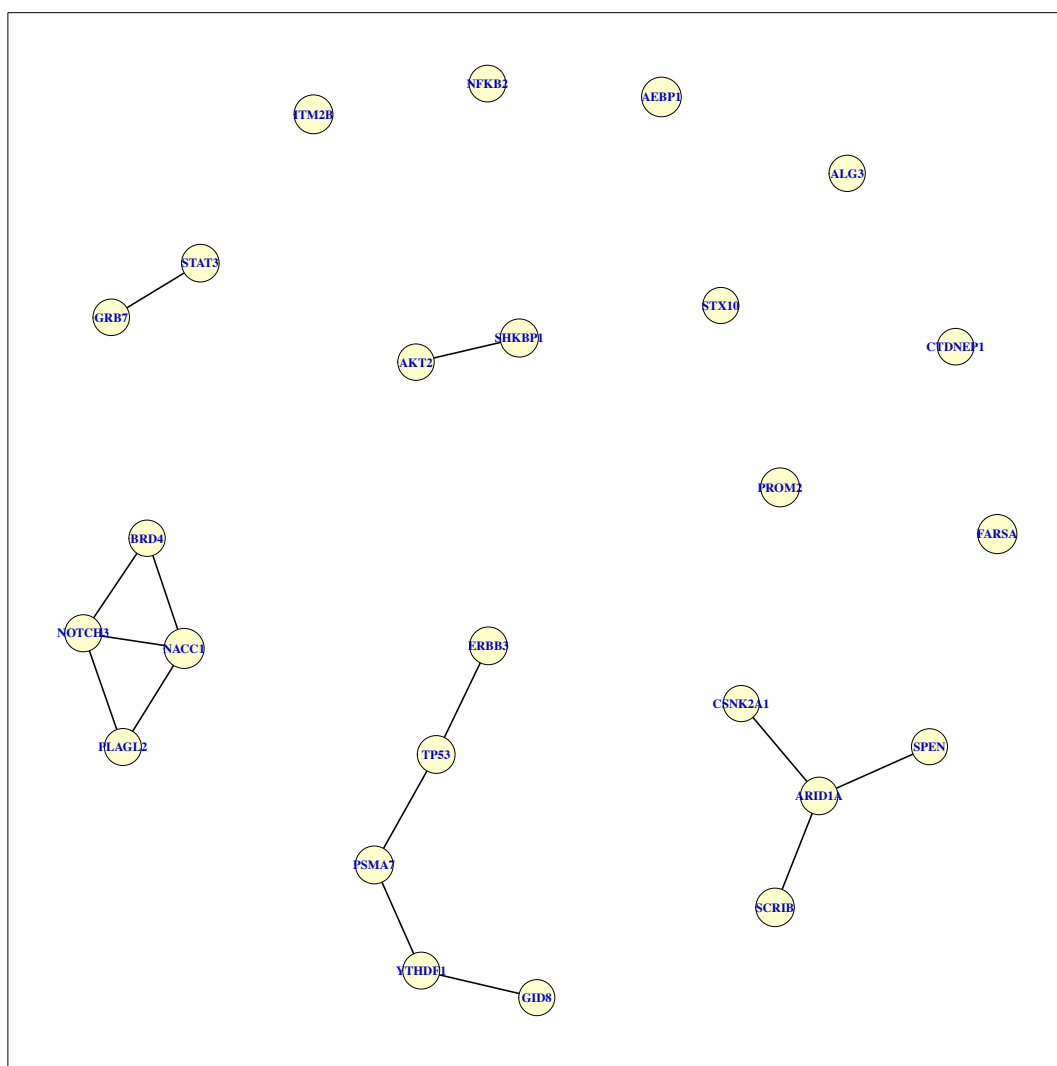


図 36 $k=3$ における遺伝子ネットワーク

$k = 3$ では NACC1, NOTCH3, BRD4 は発がんに関わる遺伝子であり, PLAGL2 はアポトーシスに関わる抑制遺伝子である. また他の ERBB3, PSMA7, SCRIB もがん遺伝子として同定され, TP3, ARID1A は有名ながん抑制遺伝子である. これらのことからがん化とその抑制に関わる遺伝子ネットワークであると解釈される.

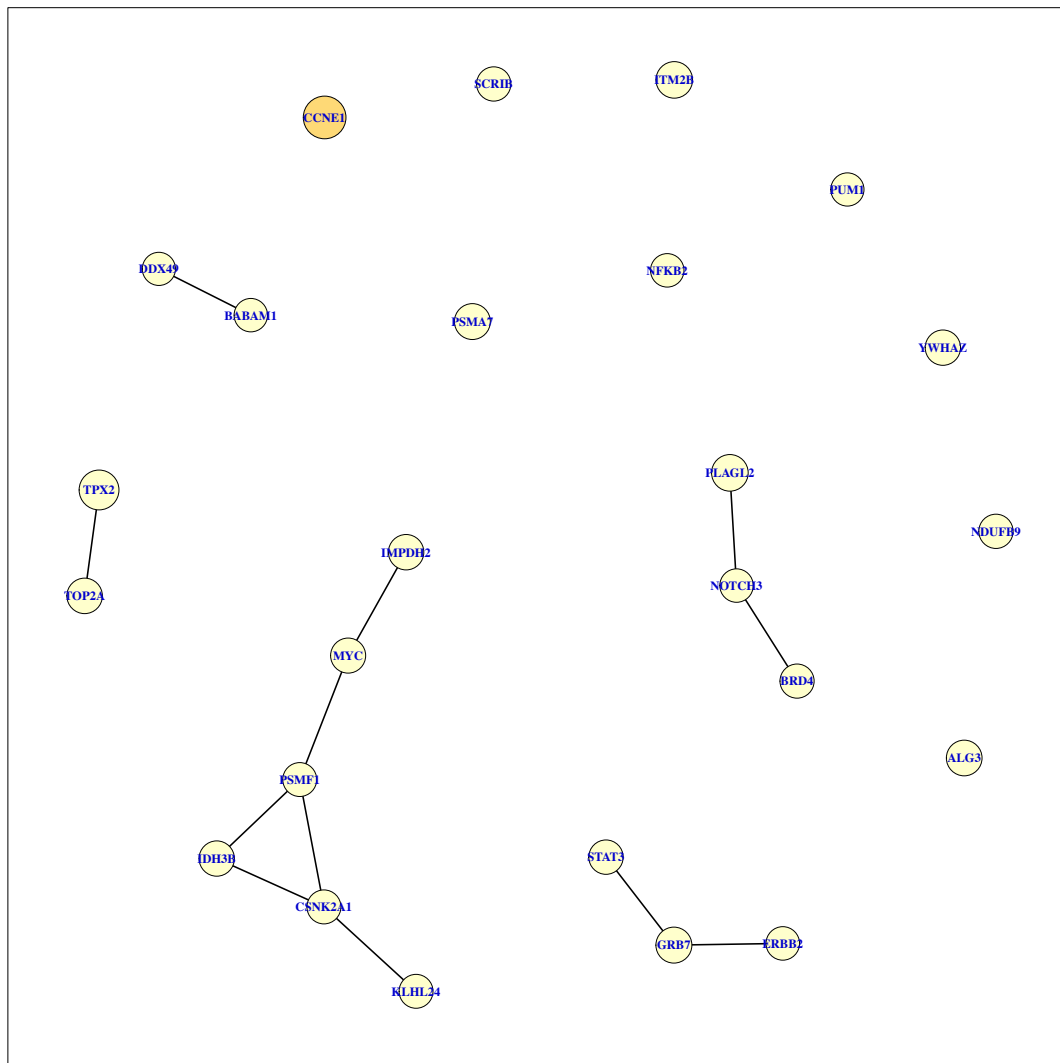


図 37 $k=21$ における遺伝子ネットワーク

$k = 21$ では CCNE1 の遺伝子と、IMPDH2, MYC, PSMF1 を含む遺伝子ネットワークが主に現れている。CCNE1 は有糸分裂の時間調整に関わるタンパク質をコードする遺伝子であり、MYC も細胞周期に関わる機能を有する。KLHL24 や CSNK2A1 はエンドサイトーシス、DNA 修復、翻訳調節、シグナル伝達、細胞周期制御、アポトーシスなどを含む様々な細胞プロセスに関与することが知られている。また、PSMF1 や IDH3B はエネルギー生成に関わる機能を持ち、IMPDH2 は DNA や RNA の合成に関与する。これらのことから、細胞分裂とそれに伴う細胞プロセスの遺伝子ネットワークであると解釈できる。

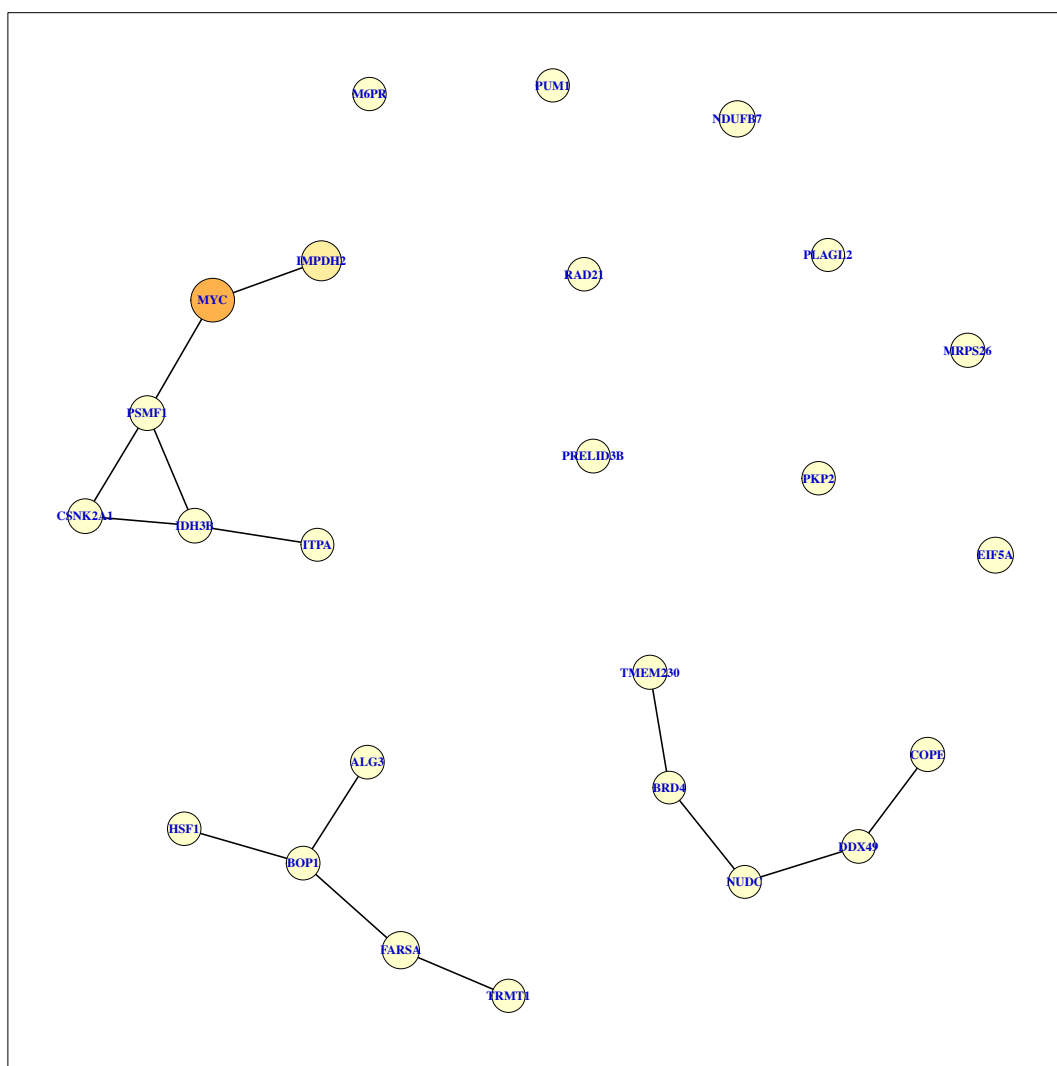


図 38 $k=41$ における遺伝子ネットワーク

$k = 41$ では MYC, IMPDH2, PSMF1 を含む遺伝子ネットワークが主に現れている． $k = 21$ とほぼ同様の遺伝子の特徴としてネットワークを形成しており，細胞周期やエネルギー活性，合成などの様々な細胞プロセスに関わる遺伝子ネットワークであると解釈できる．

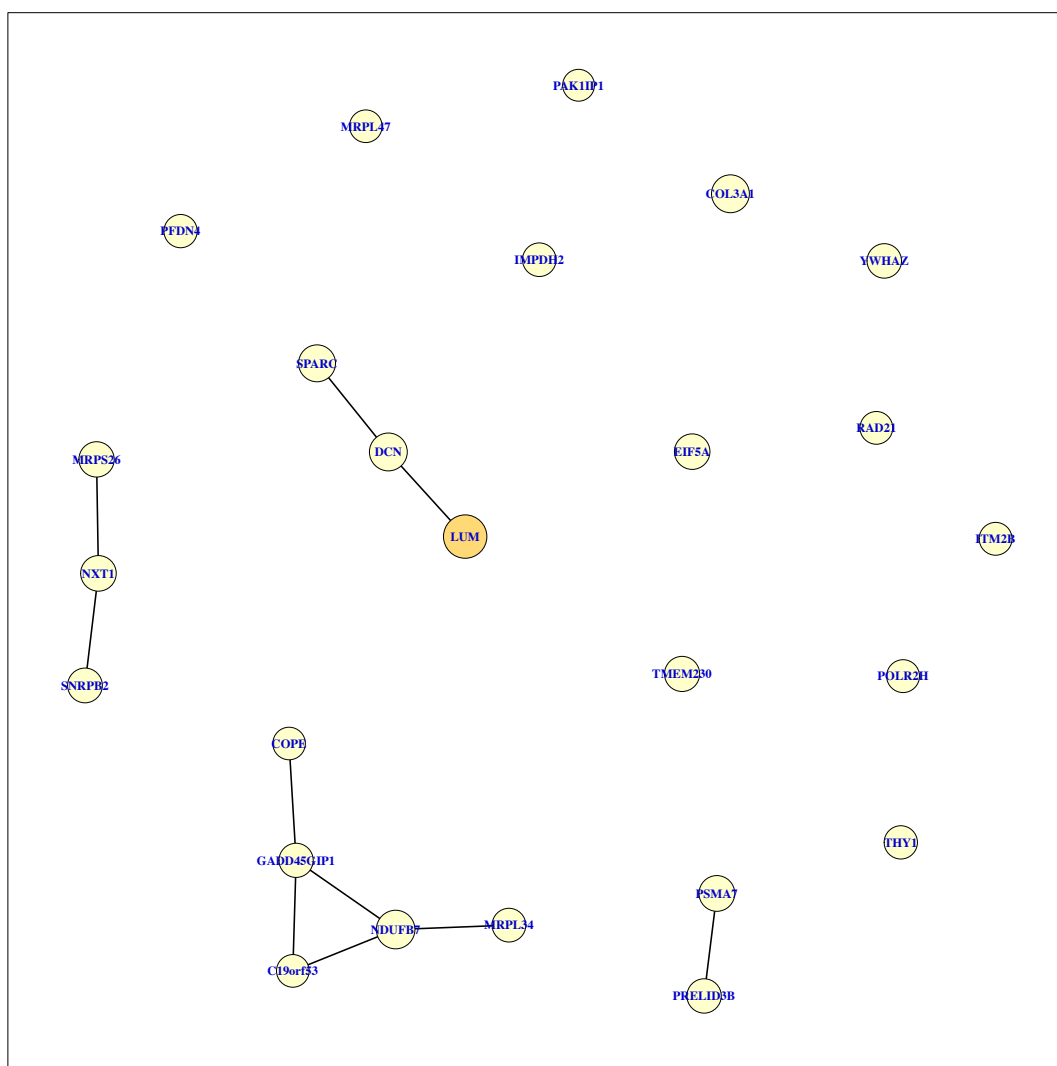


図 39 $k=30$ における遺伝子ネットワーク

$k = 30$ では LUM, DCN, SPARC が主に現れており、これらはクラスタ 1 で主に含まれるコラーゲン繊維の合成や機能に関わる遺伝子である。また、C19orf53, GADD45GIP1, COPE, NDUFB7, MRPL34 を含むネットワークも現れている。これらの遺伝子はクラスタ 2 で主に含まれており、C19orf53 は発がんなどに関わるがん遺伝子である。GADD45GIP1 はアポトーシス、COPE はゴルジ体の機能に関わる遺伝子である。NDUFB7 と MRPL34 はミトコンドリアの機能に関わる遺伝子である。

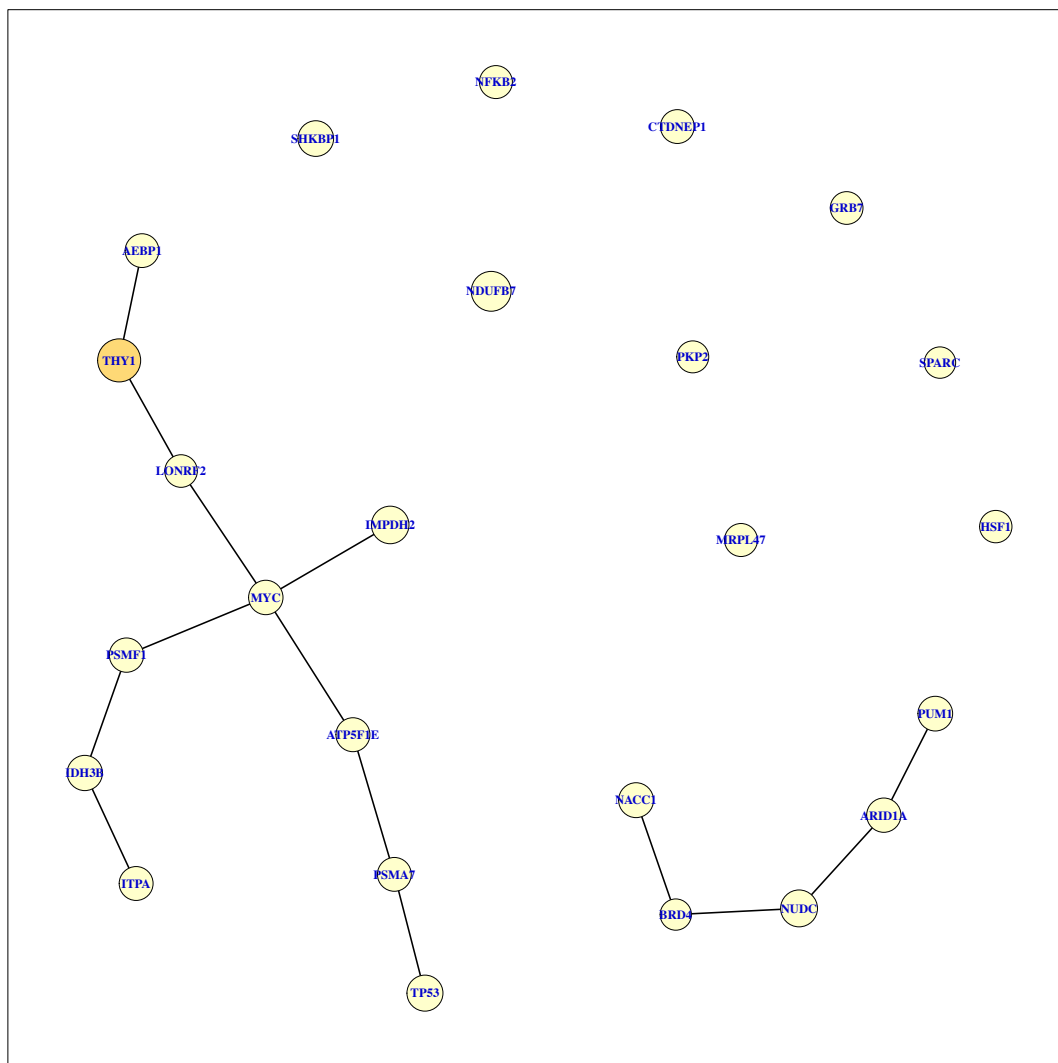


図 40 $k=16$ における遺伝子ネットワーク

$k = 16$ において、THY1 は免疫系および神経系の細胞における細胞接着および細胞伝達に関与し、LONRF2 は ATP 依存性ペプチダーゼ活性を有し、様々な細胞内機能に関連性を示すことが知られている。MYC は細胞周期に関わる機能を有し、IMPDH2 は DNA や RNA の合成に関与する。また、PSMF1 や IDH3B はエネルギー生成に関わる機能を持つ。これらの遺伝子はそれぞれが隣り合う遺伝子との機能的関連性を伺うことができる。

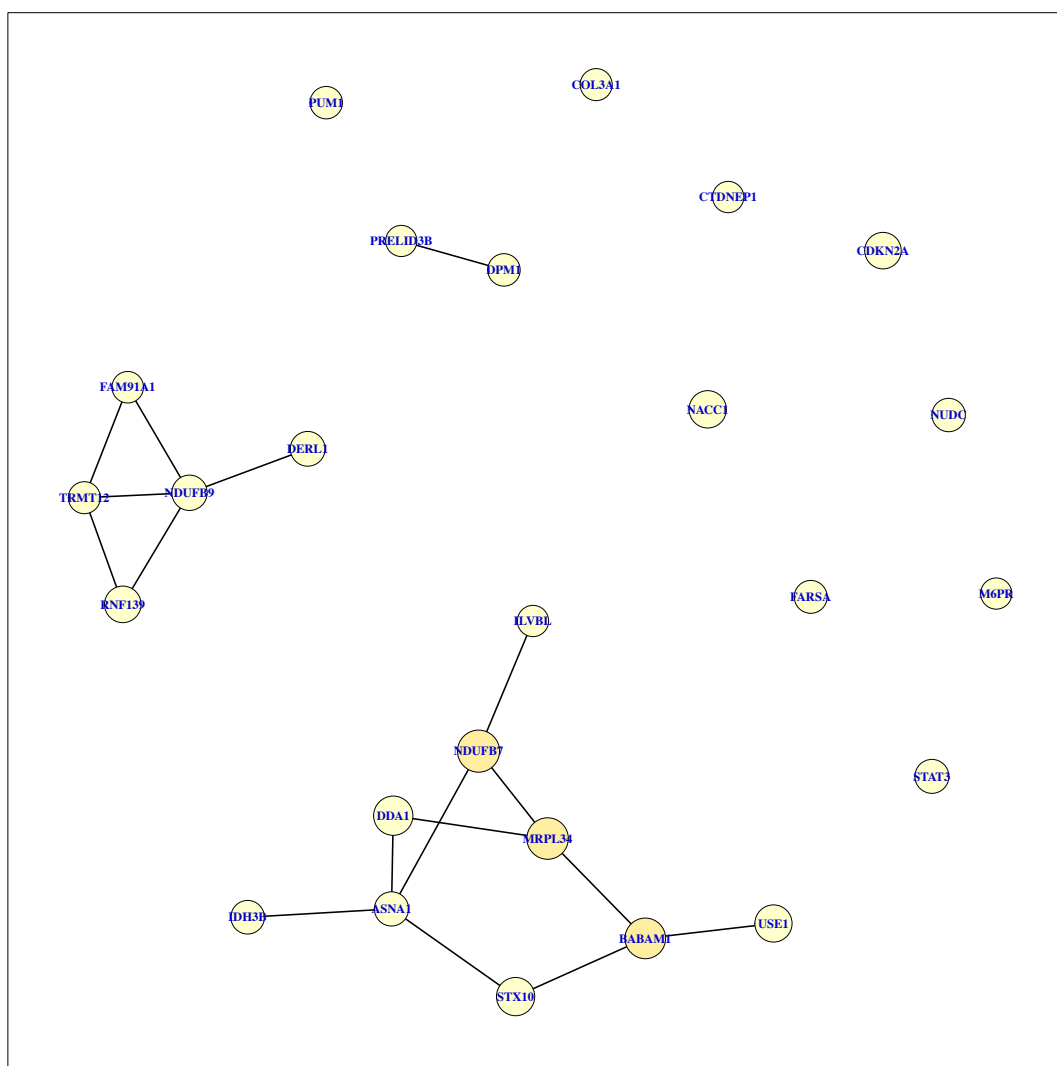


図 41 $k=22$ における遺伝子ネットワーク

$k = 22$ では $NDUFB7$, $MRPL34$, $BABAM1$ などを含む遺伝子ネットワークが主に現れている． $NDUFB7$ と $MRPL34$ はミトコンドリアの機能に関連し， $BABAM1$ は細胞周期，細胞分裂，代謝系に関わる遺伝子である．また， $ILVBL$ も代謝に関わる機能を有している． $STX10$ はゴルジ体に関連し， $ASNA1$ は小胞輸送系に関連する．これらのことから，基本的な細胞プロセスに関わる遺伝子ネットワークであると解釈できる．

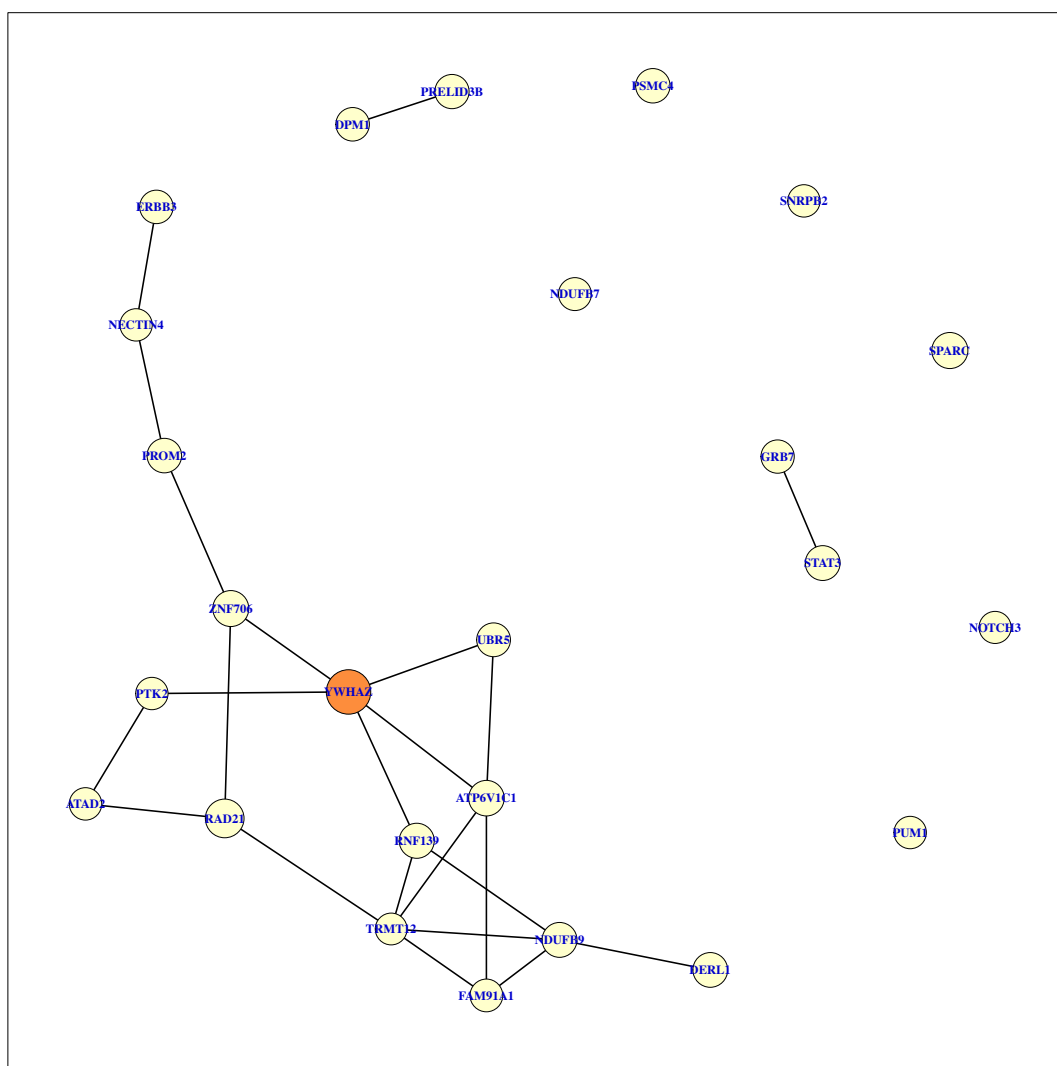


図 42 $k=37$ における遺伝子ネットワーク

$k = 37$ では YWHAZ を中心とした遺伝子ネットワークが主に現れている．YWHAZ は代謝，転写，アポトーシス，タンパク質輸送，細胞周期調節など，幅広い生物学的プロセスに関与することが知られている．また，TRMT12, UBR5, RAD21, ZNF706 などそれぞれ，リボソーム，細胞増殖・分化，コヒーシン，遺伝子発現調節に関わる遺伝子群である．これあのことから様々な細胞プロセスに関わる遺伝子ネットワークであると解釈できる．

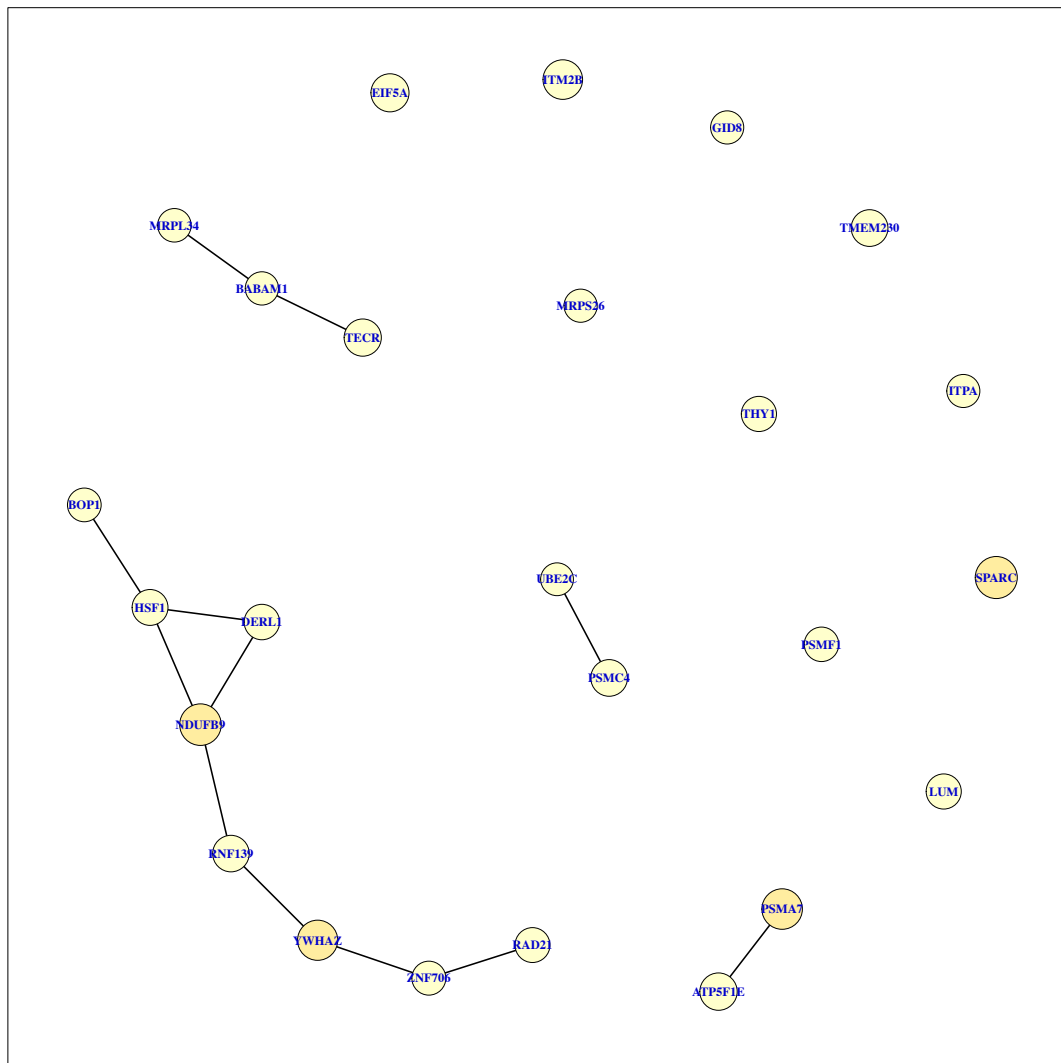


図 43 $k=4$ における遺伝子ネットワーク

$k=4$ において、YWHAZ は代謝、転写、アポトーシス、タンパク質輸送、細胞周期調節などの幅広い生物学的プロセスに関与しており、遺伝子発現制御に関わりのある ZNF706 や小胞体の機能に関わる RNFL39 とも関係がある。NDUF9 はミトコンドリアの機能に関わるタンパク質をコードしており、小胞体との関係は密接である。PSMA7 はIIの卵巣がんに伴う直腸転移が知られており、その直腸癌の肝転移と関連すると報告がある遺伝子である。卵巣がんは悪化が進むと肝臓にまで転移することが知られている。これらのことから、ミトコンドリア・小胞体などの生物学的プロセスと癌の進行を伴う遺伝子ネットワークであると考えられる。

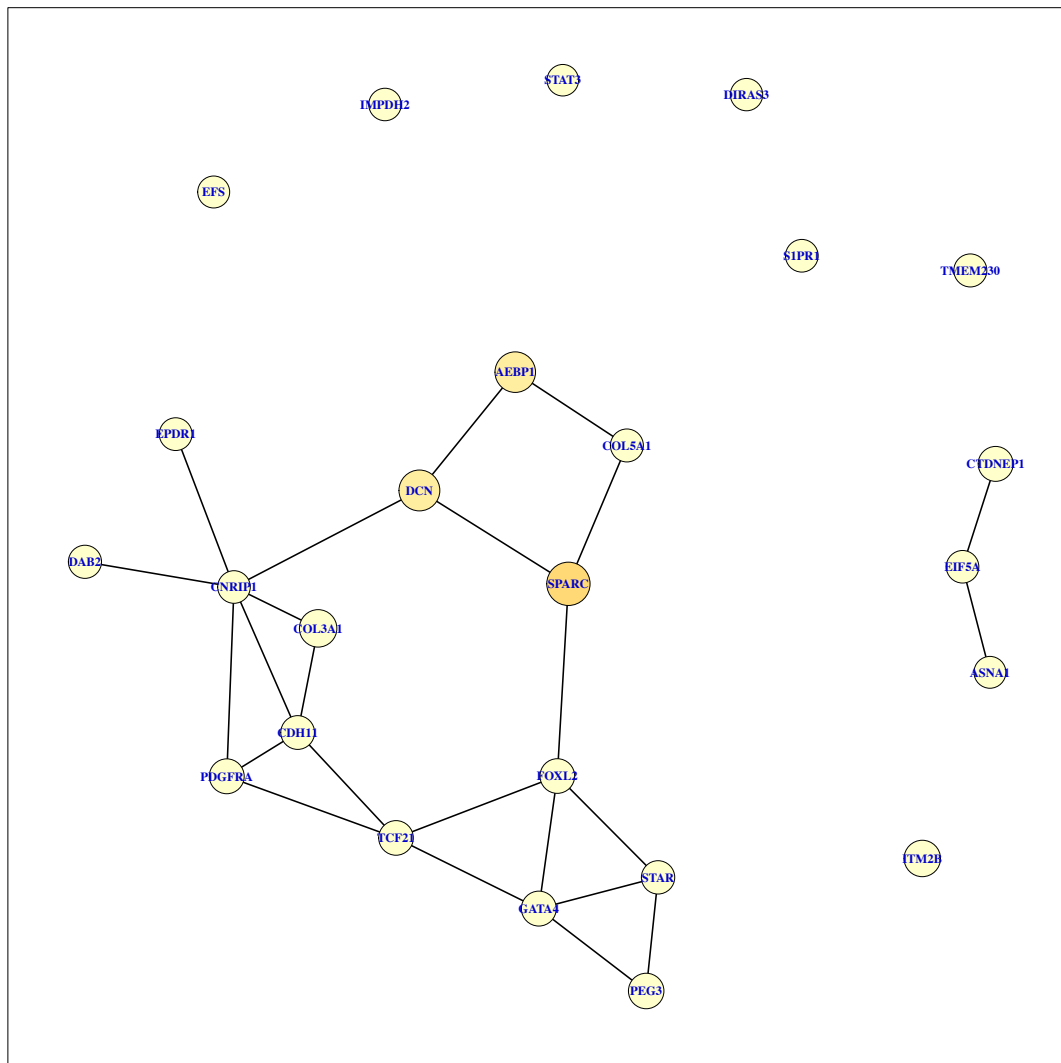


図 44 $k=40$ における遺伝子ネットワーク

$k = 40$ において、SPARC, DCN はコラーゲン繊維に関与し、AEBP1 は転写抑制因子として機能するタンパク質をコードしており、AEBP1 の遺伝子変異により不完全なコラーゲン組織を形成することが知られている。COL3A1 や COL5A1 も同様にコラーゲン繊維に関係している。一方で、FOXL2 のようなホルモン産生卵巢細胞の成長と分裂（増殖）を調節する性ホルモン系に関わる遺伝子も見受けられる。

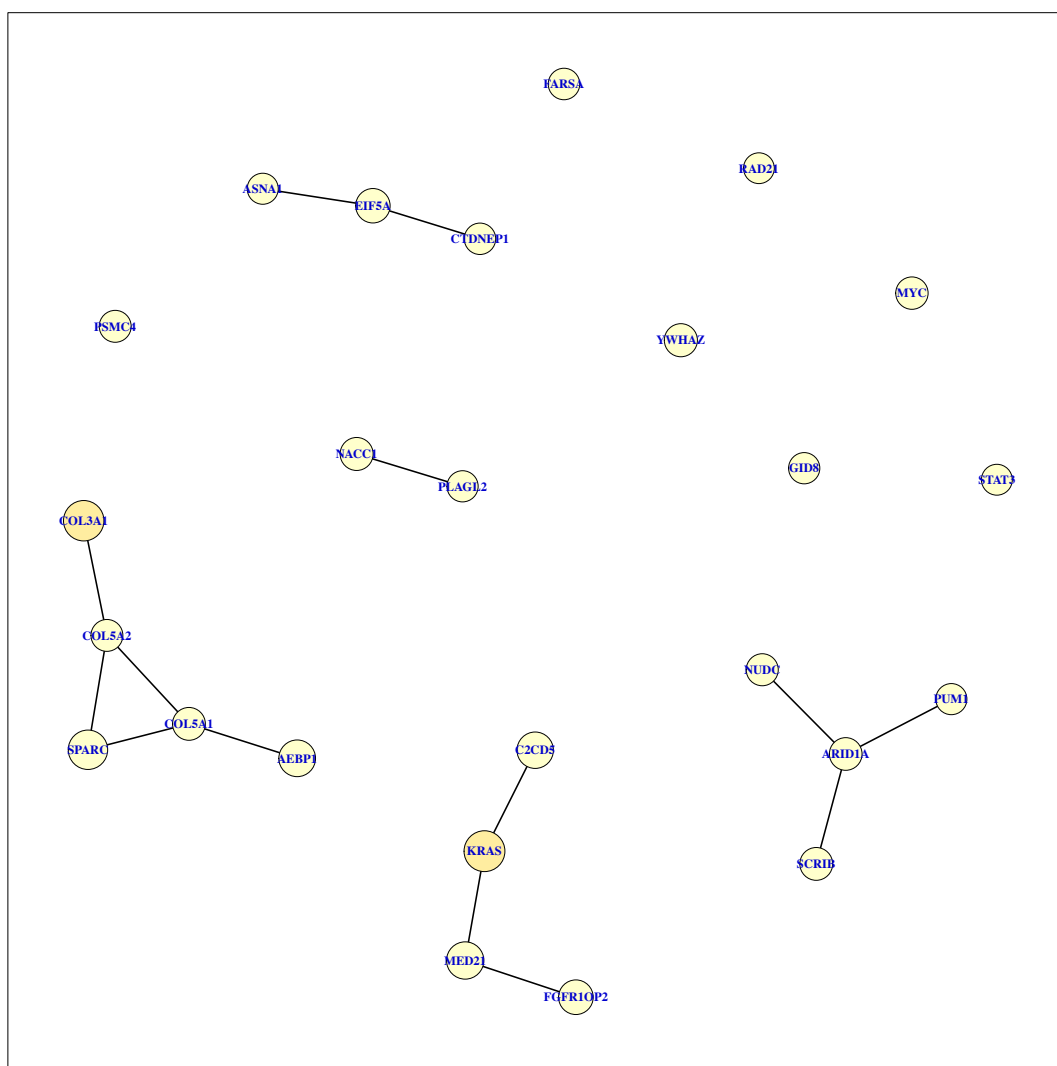


図 45 $k=2$ における遺伝子ネットワーク

$k = 2$ では COL3A1, COL5A2,SPARC を含むネットワークが特徴的に現れており、これらはコラーゲン繊維に関わる遺伝子である。また KRAS,MED21,FGFR1OP2,C2CD5 を含むネットワークも特徴的に現れている。C2CD5 はたんぱく質ドメインをコードし、KRAS のシグナル伝達に関わる機能とは関係性が深い。また KRAS は発がんの引き金となり数多くのがんに関係しており、がん細胞は RNA ポリメラーゼによる増殖や線維芽細胞の伸びた方向に浸潤していくことが知られている。

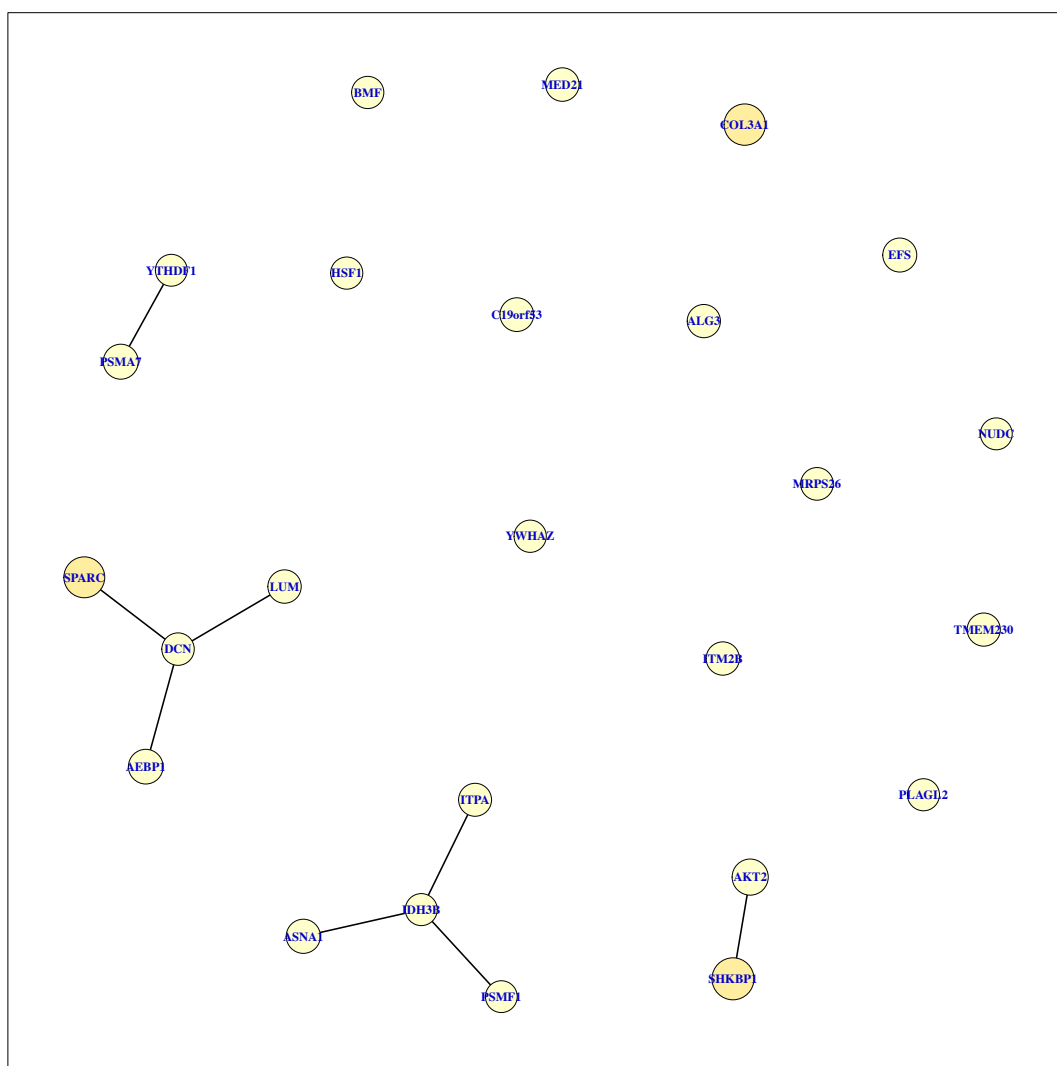


図 46 $k=5$ における遺伝子ネットワーク

$k = 5$ では COL3A1, SPARC, LUM といったコラーゲン繊維に関わる遺伝子が特徴的に現れている。また、AKT2 は卵巣がんのがん遺伝子として同定されており、SHKBF1 は子宮頸がんにおいて過剰発現するという報告がある。 $k = 40$ と同様に、コラーゲン繊維系と生殖器官に関わる遺伝子が特徴的に現れていることが分かる。

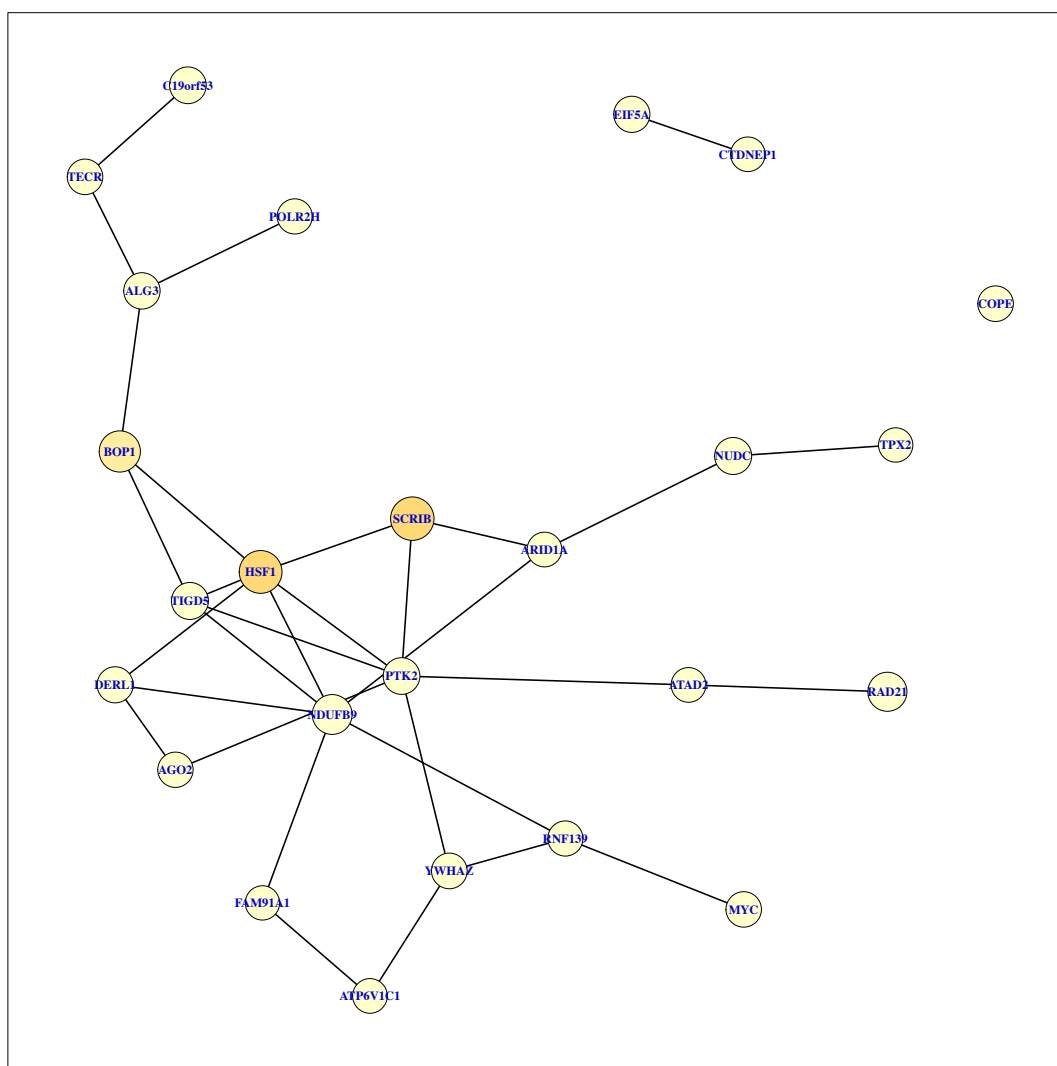


図 47 $k=9$ における遺伝子ネットワーク

$k = 9$ では SCRIB, HSF1, BOP1 などを含む遺伝子ネットワークが主に現れている. SCRIB と NUDC は細胞分裂に関与し, HSF1, BOP1, ALG3 はがん遺伝子として同定されている. また, CTDNEP1 や EIF5A も増殖に関与することから, がん細胞における増殖に関わる遺伝子ネットワークであると解釈できる.

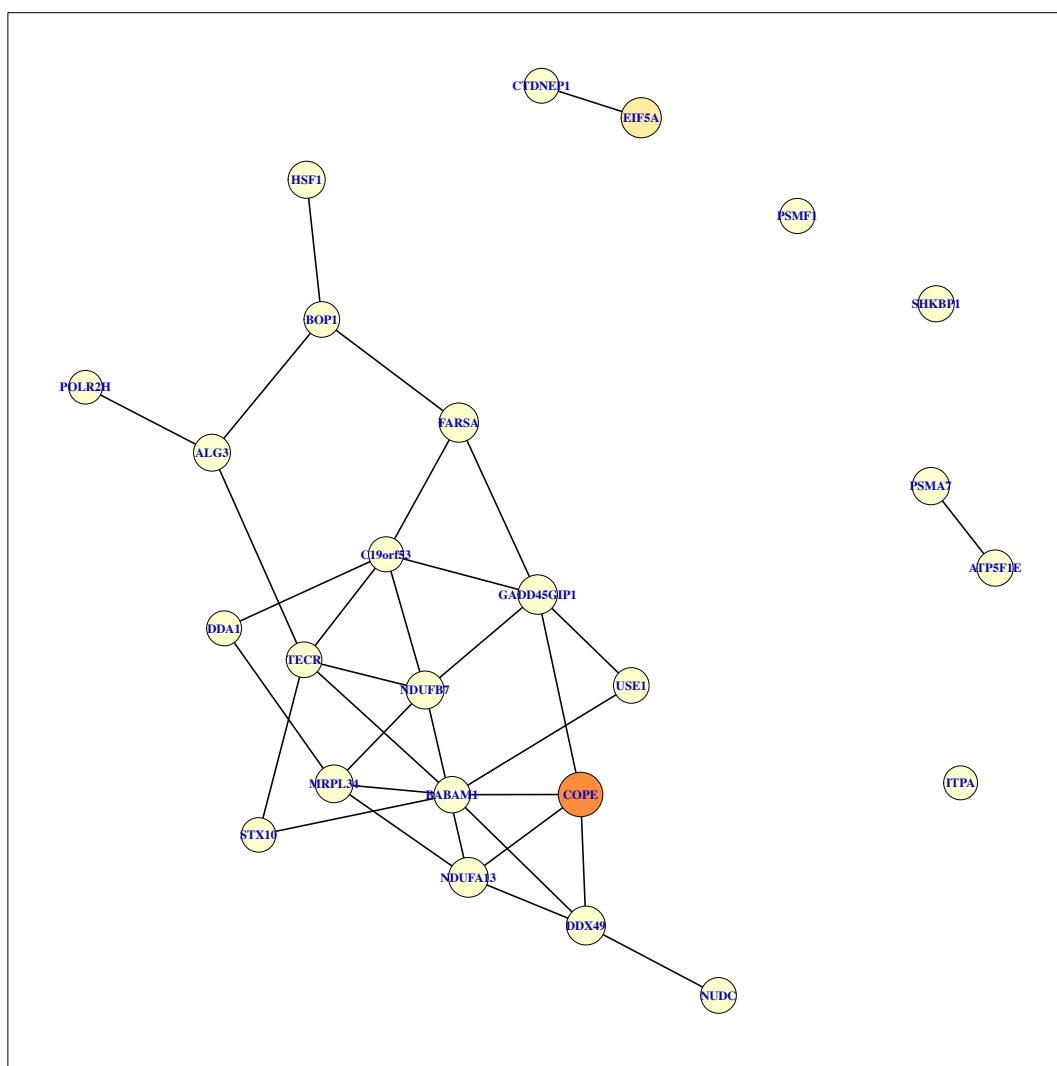


図 48 $k=14$ における遺伝子ネットワーク

$k = 14$ では COPE などを含む遺伝子ネットワークが主に現れている. COPE や TECR は小胞体輸送に関わり, HSF1, NUDC, EIF5A, CTDNEP1, DDX9 は細胞分裂・増殖に関与している. また, NDUFA13 はミトコンドリアの機能に関連すると共にアポトーシスの役割を果たすことが示唆されており, GADD45GIP1 もアポトーシスに関与している. これらのことから,

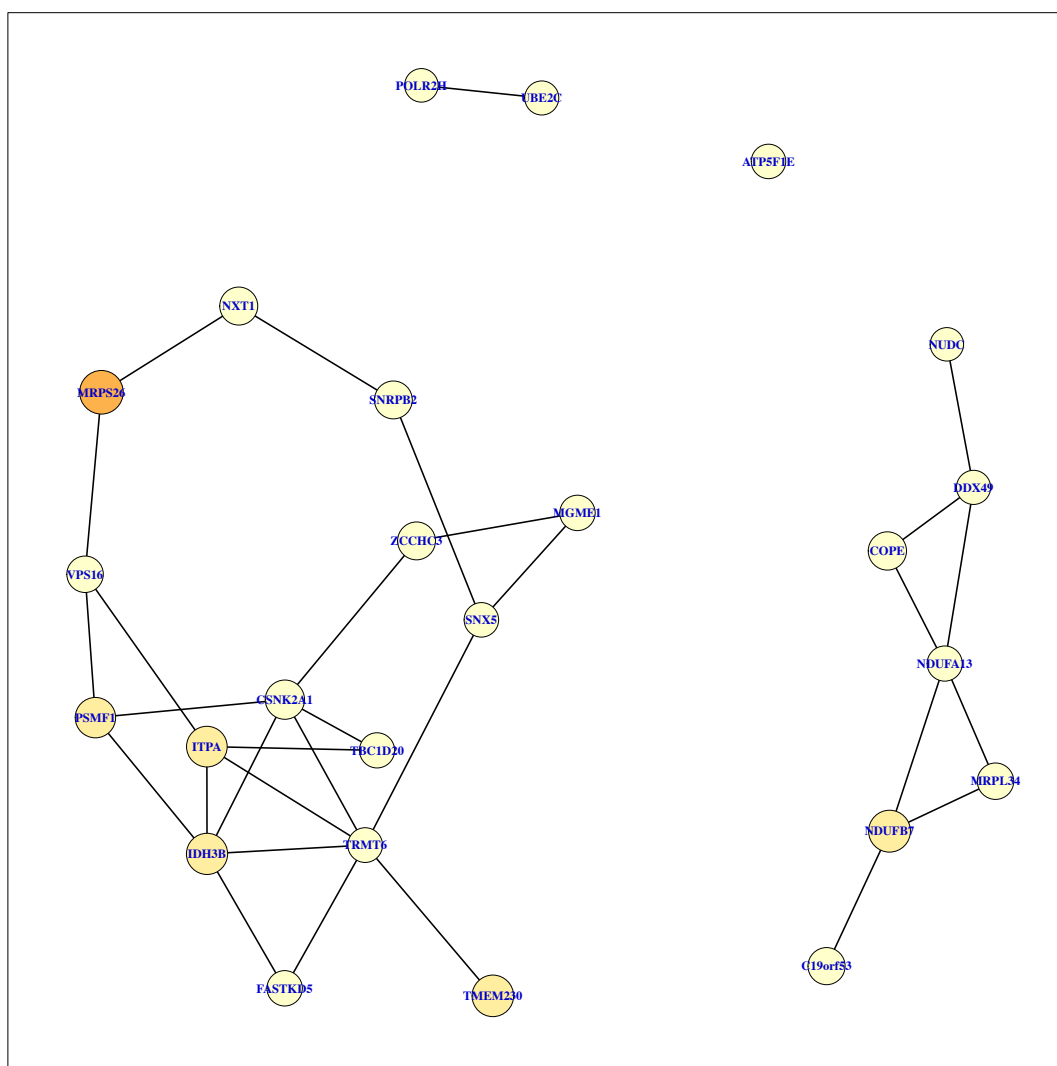


図 49 $k=10$ における遺伝子ネットワーク

$k = 10$ では MRPS26, PSMF1, NDUF7 などを含む遺伝子ネットワークが主に現れている。MRPS26 はミトコンドリア内におけるタンパク質合成に関与しており、これを中心に NXT1, VPS16, SNX5 などタンパク質輸送に関わる遺伝子がある。また、NDUF7 と共に NDUF13, MRPL34 などミトコンドリアの機能に関わる遺伝子が繋がっている。これらのことから、ミトコンドリアと小胞輸送系の遺伝子ネットワークであると解釈できる。

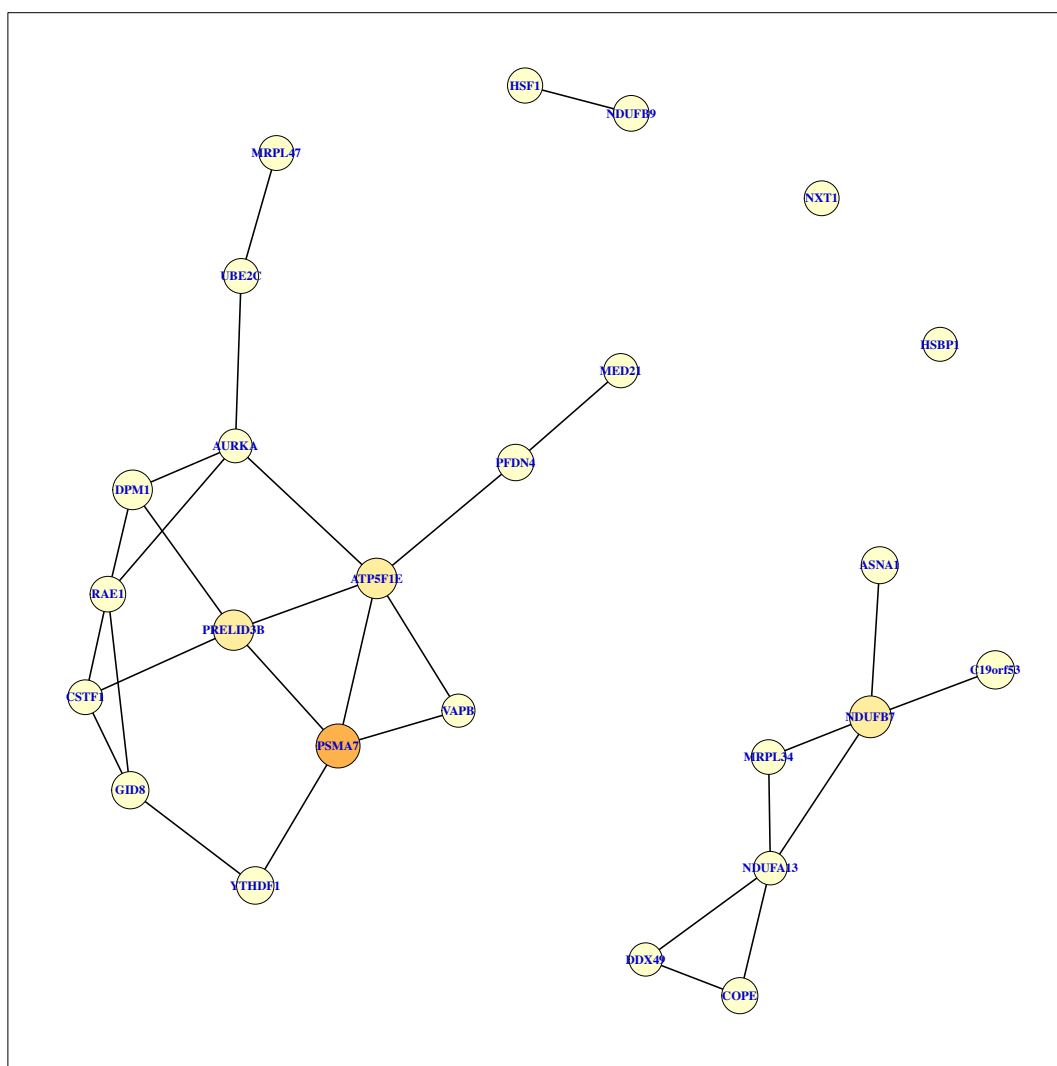


図 50 $k=18$ における遺伝子ネットワーク

$k = 18$ では PSMA7, PRELID3B, ATP5F1E, VAPB, RAE1 などを含む遺伝子ネットワークが主に現れている. PSMA7 は結腸がんの肝転移に関わる遺伝子であることが知られている. PRELID3B と ATP5F1E はミトコンドリアの機能に関与している. VAPB, RAE1 は共に小胞体輸送に関わるタンパク質をコードした遺伝子である.

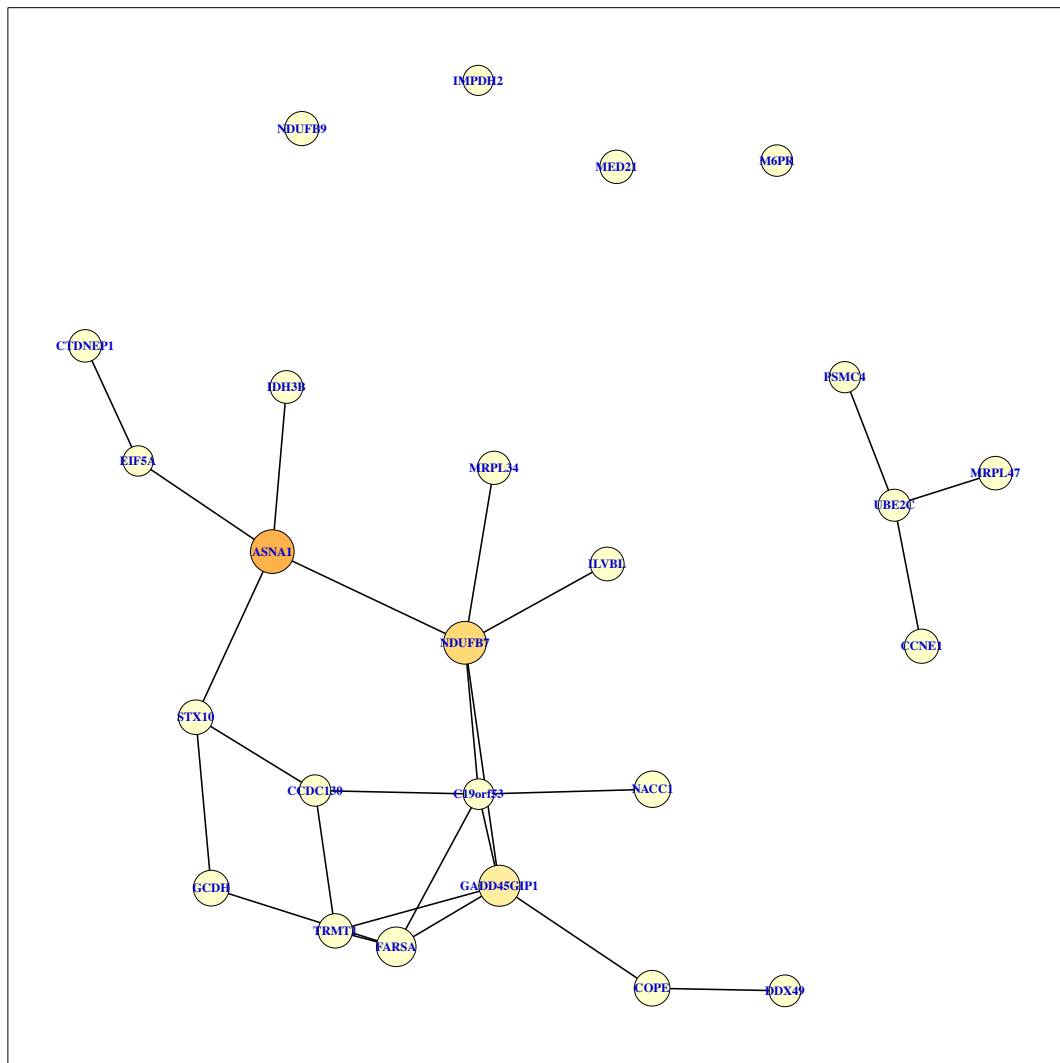


図 51 $k=7$ における遺伝子ネットワーク

$k = 7$ において、ASNA1, NDUF7, GADD45GIP1 を含む遺伝子ネットワークが主に現れている。ASNA1 はタンパク質輸送に関わるタンパク質をコードしており、NDUF7 はミトコンドリアの機能に関与している。GADD45GIP1 や NACC1 はアポトーシスに関与し、C19orf53 は卵巣がんに関わる癌遺伝子として同定されている。CCDC130 がコードするタンパク質は遺伝子発現を制御するものや転写因子など、重要な生物学的機能を持つものが多く知られている。

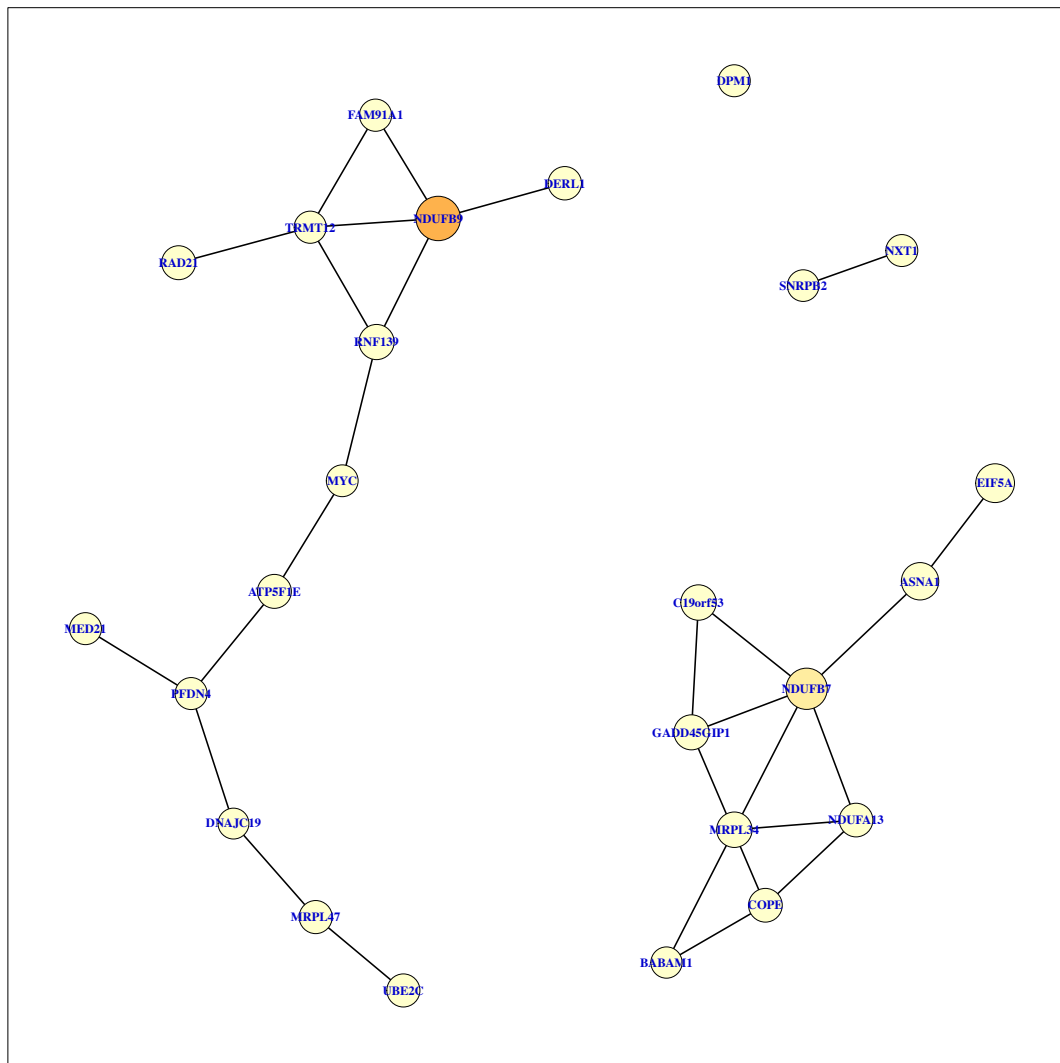


図 52 $k=15$ における遺伝子ネットワーク

$k = 15$ ではミトコンドリアの機能に関連する NDUF9 と NDUF7 を中心とするそれぞれの遺伝子ネットワークが主に現れている。NDUF13 や MRPL34 もミトコンドリア内で働くタンパク質をコードしており、GADD45GIP1, C19orf53, ASNA1 はそれぞれアポトーシス、卵巣がんの癌遺伝子、タンパク質輸送に関わるタンパク質をコードしている。

また、DERL1, RNF139 は小胞体に関係する機能を有しており、TRMT12 はリボソームに関係している。

以上のように、クラスタ 4 に分類された遺伝子ネットワークは遺伝子のもつ傾向が定まっていないものが多く、それらは基本的な細胞プロセスに関わるものだと考えられる。

第5章 考察

5.1 提案手法における特徴的な遺伝子の選択

これまでに示してきた図 12 のような遺伝子ネットワークには、エッジを持たない遺伝子がいくつか確認された。それらは基底ベクトル内でもスコアが 1 以下のような極々小さい遺伝子である。これは特徴的な遺伝子を選択する際、閾値を恣意的に決めたことにより抽出されたものと考えられる。そこで、Logistic 回帰を用いることで基底ごとに適切な閾値を設定することができ、より特徴的な遺伝子ネットワークの描画を行える可能性が考えられる。しかし、基底に含まれる要素値が二極化している場合（図 10）にうまく機能すると想定されるが、要素値が緩やかに変化する場合には特徴的な遺伝子の選択が困難であると考えられる。

5.2 遺伝子のネットワーク構造を加味した効果

同じ遺伝子発現データを用いて通常の NMF による解析も行った。NMF での基底数は BIC に基づくと 34 個であった。解析の結果、提案手法のときと同様に、それぞれコラーゲン繊維・エネルギー代謝・増殖の機能に関わる基底群が形成されていた。このことから、図 12 におけるクラスタ 1~3 で得られた基底群での特徴は、元々データがもつ特性が現れているのではないかと考えられる。一方で、NMF の解析結果の中には図 53 に示すようなネットワークも現れていた。SPARC はコラーゲン繊維に関わり、NDUFB7 はミトコンドリアの機能、TMEM230 は脳神経の機能、YWHAZ はシグナル伝達を媒介し幅広い生物学的プロセスに関係する遺伝子である。解釈という観点では、遺伝子は関連した機能を有する遺伝子と関わりを持つことから、これらのような特徴的な遺伝子が繋がりを持たず、かつ機能に纏まりが見られないことは、生物学的知見の獲得やモデル構築が困難になると考えられる。このことから、遺伝子ネットワークを加味した提案手法の方が解釈上有利に働くと考えられる。

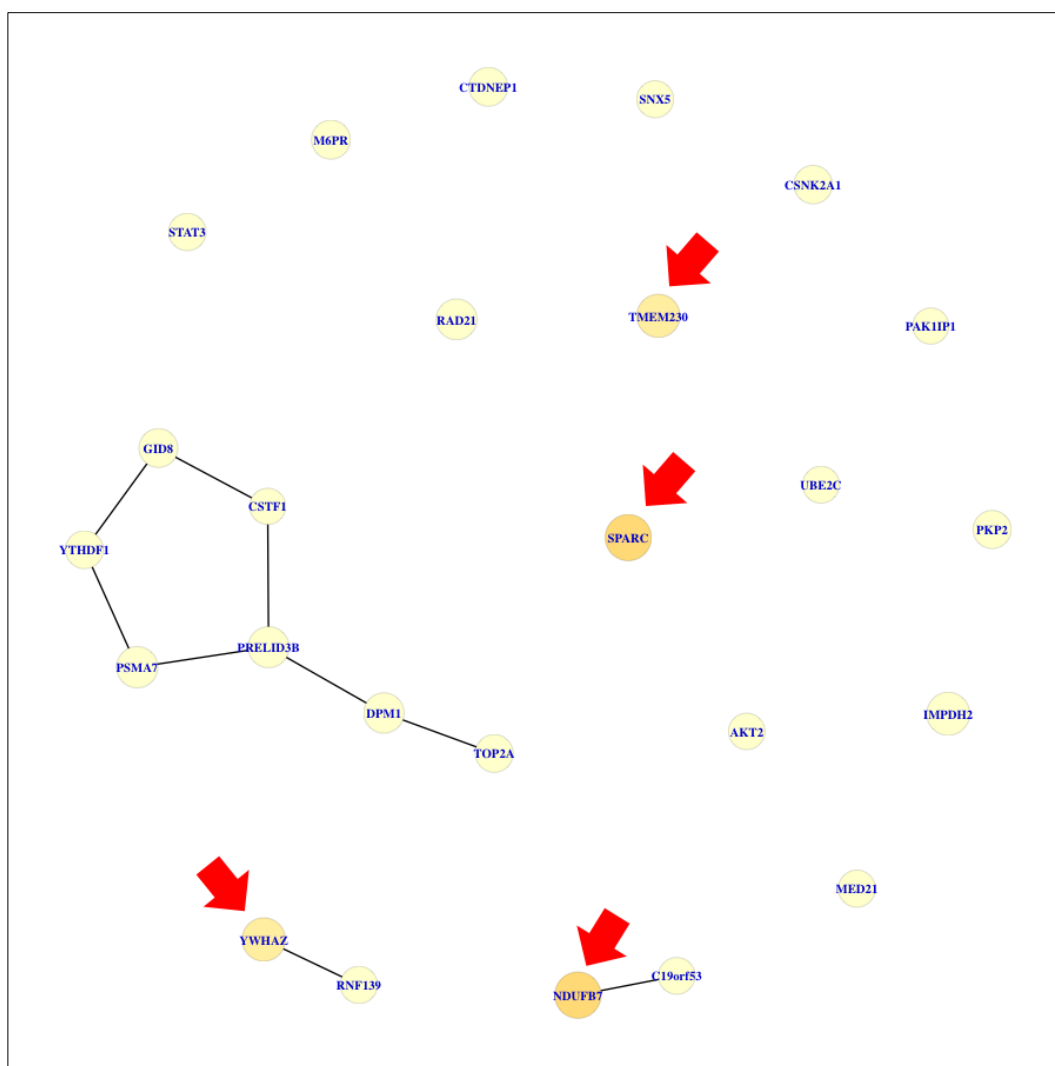


図 53 NMF での解析結果の一例

5.3 各基底における臨床情報について

行列 V にはそれぞれの患者の各基底に対する重みが格納されている。そこで、それぞれの基底ベクトルの上位 10% の値をもつ患者の臨床情報を得ることで、各基底に対する臨床情報の特性を見ることができる。臨床情報として年齢に着目したが、それぞれの基底で特徴的な臨床情報は得られなかった。これには 2 つの理由が考えられる。1 つ目は、TCGA データベースや卵巣がんによるデータの偏りである。他の臨床情報として、人種・性別・がんのステージなどの情報が含まれていたが、人種はほぼ全てが白人であった。用いているデータが卵巣がんであるため性別は全てが女性であり、自覚症状のないがんであるため、ステージはほぼ全てがⅢかⅣであった。このことから、データそ

のものに差が生じていないことが挙げられる。このような問題に対し、他のデータベースの利用や他種のがんデータを用いることで、偏りの緩和・他のクリニカル情報に関する検討を行うことができる可能性がある。2つ目は、提案手法に患者に対する制約条件がなかったことが挙げられる。提案手法では、遺伝子ネットワークの構造に対してのみ制約条件が加えられていたため、患者間の遺伝子発現パターンなどの類似度を考慮する制約条件を加えることで、クリニカル情報においても特徴を得られる可能性が考えられる。

5.4 AIC による基底数の推定

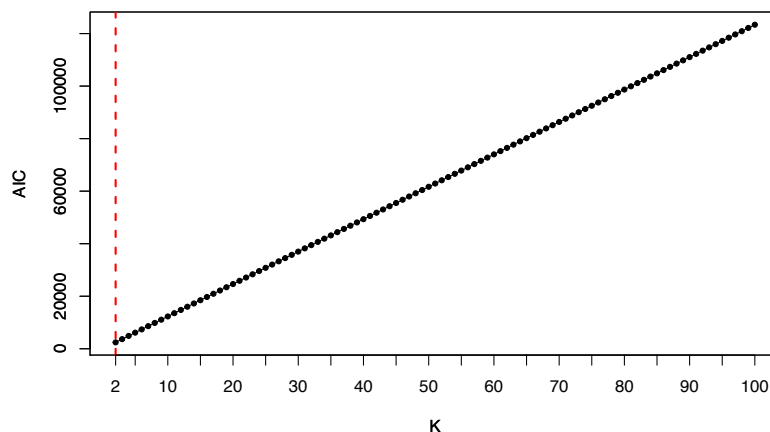
本研究では BIC に基づいた基底数で解析を行ったが、モデルによる汎化誤差を評価する AIC を用いた際の挙動も確認した。AIC の定義に従うと次のように表される [22]。

$$AIC = -2\log(D_{EU}) + 2k(i + j) \quad (5.1)$$

また、他の論文[23]では以下のような式も定義されていた。

$$AIC = -\log(D_{EU}) + (ik + kj - k^2) \quad (5.2)$$

(5.1)(5.2)式を用いた際の基底数 k による AIC の遷移を以下に示す。図中の赤点線は AIC が最小であった時の基底数 k を示している。



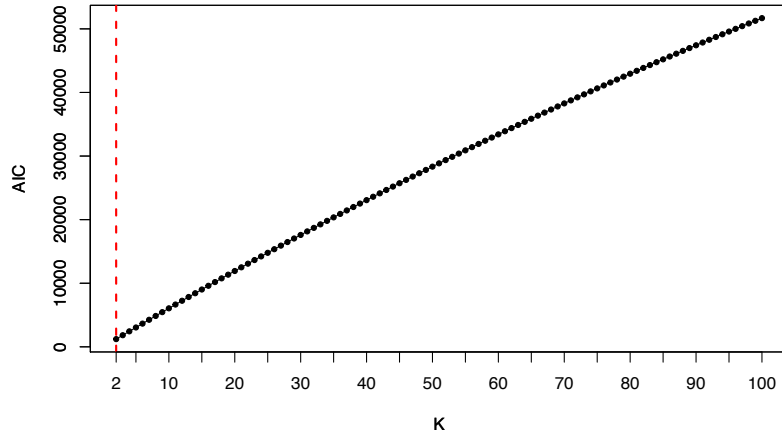


図 54 基底数 k に対する AIC の遷移 (上 : 5.1 式, 下 : 5.2 式)

BIC での基底数は 41 だったのに対し, AIC では最適な基底数が 2 と推定された. この基底数では大規模ネットワークから 2 つの遺伝子ネットワークを描画することしができず, 解釈に貢献し難いと言える. また, NMF では分解後の行列 UV に含まれる要素全てがパラメータである. 本研究で用いた入力データは 245×374 の行列であったため, 基底を 1 つ増やした際にパラメータが $245 \times 1 + 1 \times 374 = 619$ 個増加することになる. パラメータの増加に対して誤差の減少が少ないために, AIC の第一項よりも第二項が著しく大きくなるため, AIC が図 35 のような単調増加になったと考えられる. ただし, NMF の基底数の推定に AIC を使っている論文も見受けられる [22][23]. これはデータの次元が低い際に, 誤差の減少量に対して基底数の増加に伴うパラメータの増加量が低いために上手く機能すると考えられ, 本研究のような高次元データに対しては BIC が適切であったと言える.

5.5 λ の値による基底数の変動

提案手法において, 以下のような最適化問題を設定している.

$$\begin{aligned} & \text{minimize} \quad \|X - UV^T\|_F^2 + \lambda \text{tr}(U^T L U) \\ & \text{subject to} \quad U \geq 0, V \geq 0 \end{aligned}$$

解析ではパラメータ $\lambda = 1$ と設定し行なっているが, λ の値を変化させた際の挙動を確認

した。 $\lambda = 0.1, 1, 10, 50, 100$ と変化させ、それぞれで基底数を変化させた時の誤差と BIC の遷移を以下に示す。図 56 中において、BIC が最小であった基底数を点で示している。

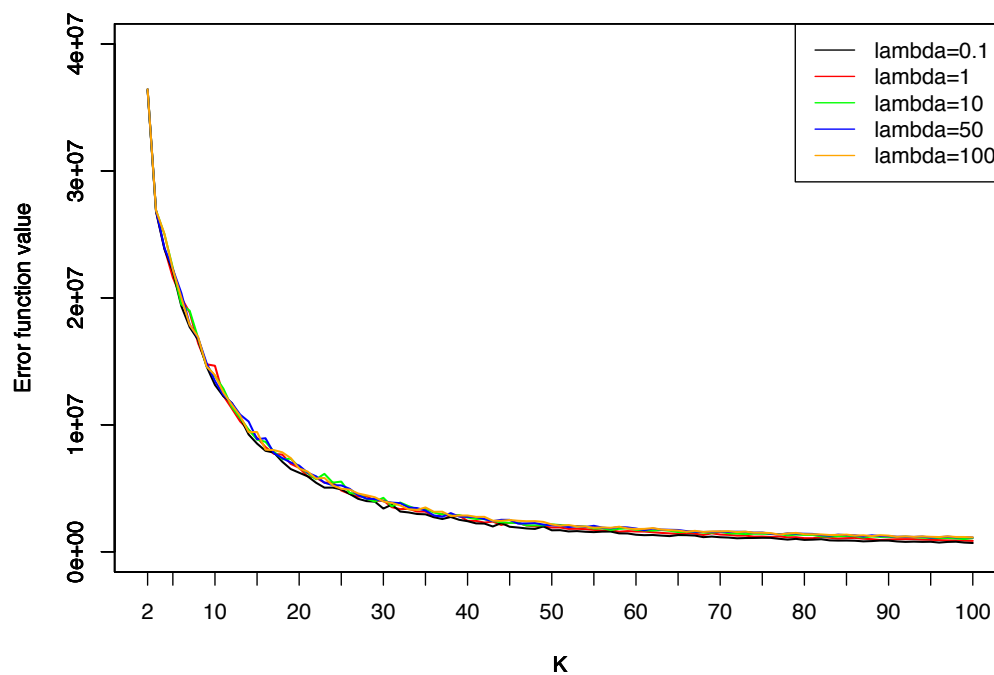


図 55 各 λ による基底数 k に対する誤差の遷移

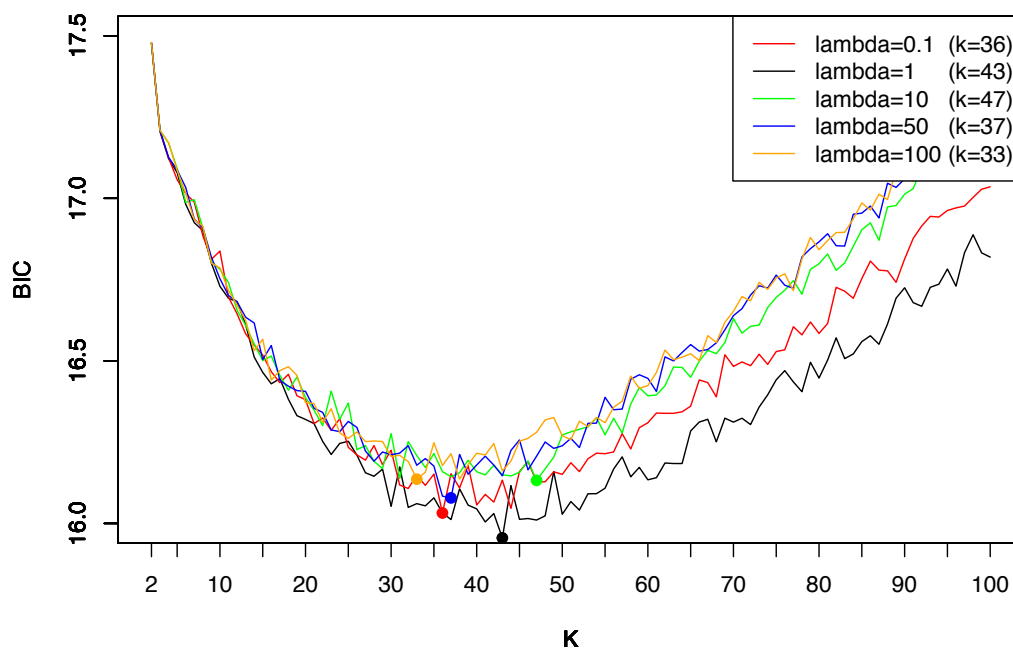


図 56 各 λ による基底数 k に対する BIC の遷移

上記の図より、 λ の値を変化させた際に誤差、BIC共に大きな変化は見られなかった。これは提案手法での理論上、行列 U 内において結合を持つ遺伝子同士は似た値に近づくことで誤差を小さくするため、 λ の値を10,100,...と変化させても基底数 k に直接影響はないと考えられる。そのため、図56中での各 λ におけるBICの最小値の違いは、要素更新時の乱数による誤差によるものであると考えられる。

また、各 λ での遺伝子ネットワークの描画についても、全く同じネットワークではないものの、最も特徴的である遺伝子をネットワークに含むことが確認された。

これらのことから、 λ の値による結果への特徴的な変化が見受けられないため、 λ への入力値の上限を1とするような条件式、あるいは指数関数などを用いることでパラメータ値の自由度を抑えるように設定することも考えられる。

第6章 結論

大規模で複雑な構造遺伝子ネットワークから生物学的知見の獲得やモデルの構築を行う場合、一般的に小規模なネットワークを形成することから行う。しかし、この時、遺伝子発現データあるいは遺伝子ネットワークの持つグラフ構造のどちらか単一の情報にのみ基づいて行われる。そこで、本研究では、より良い解釈が行われるために両データの情報を加味した手法を提案し、得られた小規模な遺伝子ネットワークの解釈を行った。

提案する手法では、非負値行列因子分解を基盤手法とし、グラフ構造を加味できるようにラプラシアン行列を組み込んだ更新式を導出した。また、ベイズ情報基準量に基づいた 41 個の基底で解析を行い、それぞれの基底で上位 10% のスコアをもつ遺伝子を特徴的なものと暫定し、それらの遺伝子から成るネットワークの解釈を行った。

41 個の基底は階層的クラスタリングから大きく 4 つに分類され、そのうち 3 つのクラスターでは、それぞれ、コラーゲン繊維・エネルギー代謝・細胞分裂に関する機能を有する遺伝子で構成されたネットワークが現れていた。これに加え、それぞれの基底では他の機能に関する遺伝子で構成されたネットワークが現れていることが分かった。また、同じデータを用いた通常非負値行列因子分解での解析では、特徴的な遺伝子がネットワークを持たず、また機能的に纏まりが見られない結果が得られた。このことから、ネットワーク構造を加味した提案手法では、より良い解釈を行える結果を得たことを示した。これらのことから、細胞内の活動におけるシステムの連動性について伺うことができ、より良い生物学的知見の獲得やモデルの構築に貢献する可能性を示唆した。

謝辞

本研究は多くの方々からの御指導，御協力をいただき，進めることができました．本研究を行うにあたり，丁寧な御指導と研究環境を与えてくださった奈良先端科学技術大学院大学情報科学研究科計算システムズ生物学的研究室 金谷 重彦 教授に深く感謝致します．また，研究室ゼミでは的確な御意見を頂きました情報科学研究科 小野 直亮 准教授，情報科学研究科 Md.Ataf-UI-Amin 准教授，情報科学研究科 黄 銘 助教に厚く御礼申し上げます．計算システムズ生物学的研究室内の皆様には，大学院生活全般にわたり様々な面で支えていただきました．皆様のおかげで充実した大学院生活を過ごすことができました．誠にありがとうございました．

最後に，大学院修了まで多大なる支援を頂いた家族に深く感謝し，ここに謝辞とさせていただきます．

参考文献

- [1] “人口動態統計月報年計（概数）の概状”，平成 29 年度版，厚生労働省，<
<https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai17/dl/gaikyou29.pdf>
>
- [2] 日本バイオインフォマティクス学会，“バイオインフォマティクス入門”，慶應義塾大学
出版会, pp.154-159, (2015)
- [3] Lee W. P., Tzou W. S. “Computational methhods for discovering gene networks from
expression data”, Brief. Bioinform. **10**, pp.408-423, (2009)
- [4] He F., Balling R., Zeng AP., “Reverse engineering and verification of gene networks:
principles, assumptions, and limitations of present methods and future perspectives.”,
J.Biotechnol., **144**, pp.190-203, (2009)
- [5] Joshua M. Stuart, Eran Segal, Daphne Koller, Stuart K. Kim, “A Gene-Coexpression
Network for Global Discovery of Conserved Genetic Modules”, Science, **302**, pp.249-
255, (2003)
- [6] Insuk Lee, Shailesh V. Date, Alex T. Adai, Edward M. Marcotte, “A Probabilistic
Functional Network of Yeast Genes”, Science, **306**, pp.1555-1558, (2004)
- [7] Friedman N., Linial M., Nachman I., Pe'er D., “Using Bayesian networks to analyze
expression data.”, J. Comput Biol., **7(3-4)**, pp.601-620, (2000)
- [8] Imoto S., Kim S., Goto T., Miyano S., Aburatani S., Tashiro K., Kuhara S., “Bayesian
network and nonparametric heteroscedastic regression for nonlinear modeling of
genetic network.”, J. Bioinform Comput Biol., **1(2)**, pp.231-252, (2003)
- [9] Muna Affara, Benjamin Dunmore, Christopher Savoie, Seiya Imoto, Yoshinori Tamada,
Hiromitsu Araki, D. Stephen Charnock-Jones, Satoru Miyano, “Understanding
endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal?”,
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B: BIOLOGICAL
SCIENCES, **362**, pp.1469-1487, (2007)
- [10] Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera
& Andrea Califano, “Reverse engineering of regulatory networks in human B cells”,

- Nature Genetics, **37**, pp.382–390, (2005)
- [11] Qingyang Zhang, Joanna E Burdette, Ji-Ping Wang, “Integrative network analysis of TCGA data for ovarian cancer”, BMC Systems Biology, (2014)
 - [12] Md. Altaf-Ul-Amin, Hisashi Tsuji, Ken Kurokawa, Hiroko Asahi, Yoko Shinbo, Shigehiko Kanaya, “DPClus: A Density-periphery Based Graph Clustering Software Mainly Focused on Detection of Protein Complexes in Interaction Networks”, Journal of Computer Aided Chemistry , **7**, pp.150-156, (2006)
 - [13] Lee D.D., Seung H.S., “Learning the parts of objects by non-negative matrix factorization”, Nature, **401**, pp.788- 791, (1999)
 - [14] Paatero P., U. Tapper, “Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values”, Environmetrics, **5**, pp.111-126, (1994)
 - [15] 亀岡弘和, "計測・センシングのアルゴリズム, 非負値行列因子分解", (計測と制御 Vol.51. No.09), 計測自動制御学会, p.p.835-844
 - [16] 李莉, 亀岡弘和, 牧野昭二, "補助関数法による識別的 NMF の基底学習アルゴリズム", 日本音響学会講演論文集, p.p.519-522, (2017)
 - [17] 三宅敏恒, "入門微分積分", 培風館, (1992)
 - [18] 清水昭比古, "学力低下時代の教え方, 第 4 回, ラグランジの未定係数法", (日本機械学会誌 第 112 巻 第 1093 号)、一般社団法人日本機械学会, p.p.987-992, (2009)
 - [19] D. A. Spielman, "Spectral graph theory", Chapter 18 of Combinatorial Scientific Computing (Eds. U. Naumann & O. Schenk), p.p.495-524, (2012)
 - [20] S.Boyd, L.WNDENBERGHE, “Convex Optimization”, Cambridge Universit Pyress, (2004)
 - [21] Owen, A.B. & Perry, P.O., “Bi-cross-validation of the SVD and the nonnegative matrix factorization”, The Annals of Applied Statistics, 3(2), pp.564–594, (2009)
 - [22] Kenji Watanabe, Akinori Hidaka, Nobuyuki Otsu, Takio Kurita, “Automatic Analysis of Composite Physical Signals Using Non-Negative Factorization and Information Criterion”, PLos One, (2012)

- [23] Sayaka Yamauchi, Masanori Kawakita, Junichi Takeuchi, “Botnet Detection Based on Non-negative Matrix Factorization and the MDL Principle” International Conference on Neural Information Processing, pp.400-409, (2012)

付録

表 1 解析に用いた遺伝子一覧

Ensemble ID	Symbol	Gene Name
ENSG00000139687	RB1	Retinoblastoma 1
ENSG00000171862	PTEN	Phosphatase And Tensin Homolog
ENSG00000153071	DAB2	Disabled Homolog 2, Mitogen-Responsive Phosphoprotein
ENSG00000008226	DLEC1	Deleted In Lung And Esophageal Cancer 1
ENSG00000141510	TP53	Tumor Protein P53
ENSG00000196712	NF1	Neurofibromin 1
ENSG00000113140	SPARC	Secreted Protein, Acidic, Cysteine-Rich
ENSG00000184012	TMPRSS2	Transmembrane Protease, Serine 2
ENSG00000132906	CASP9	Caspase 9, Apoptosis-Related Cysteine Peptidase
ENSG00000118495	PLAGL1	Pleiomorphic Adenoma Gene-Like 1
ENSG00000186153	WWOX	WW Domain Containing Oxidoreductase
ENSG00000071242	RPS6KA2	Ribosomal Protein S6 Kinase, 90kDa, Polypeptide 2
ENSG00000012048	BRCA1	Breast Cancer 1, Early Onset
ENSG00000139618	BRCA2	Breast Cancer 2, Early Onset
ENSG00000162595	DIRAS3	DIRAS Family, GTP-Binding RAS-Like 3
ENSG00000198300	PEG3	Paternally Expressed 3
ENSG00000117713	ARID1A	AT Rich Interactive Domain 1A
ENSG00000183715	OPCML	Opioid Binding Protein/Cell Adhesion
ENSG00000065559	MAP2K4	Mitogen-Activated Protein Kinase Kinase 4
ENSG00000155495	MAGEC1	Melanoma Antigen Family C, 1
ENSG00000060709	RIMBP2	RIMS Binding Protein 2
ENSG00000136997	MYC	V-Myc Myelocytomatosis Viral Oncogene Homolog
ENSG00000164045	CDC25A	Cell Division Cycle 25A
ENSG00000121879	PIK3CA	Phosphatidylinositol-4, 5-Bisphosphate 3-Kinase
ENSG00000074181	NOTCH3	Notch 3
ENSG00000163577	EIF5A2	Eukaryotic Translation Initiation Factor 5A2
ENSG00000168610	STAT3	Signal Transducer And Activator Of Transcription 3
ENSG00000139083	ETV6	Ets Variant 6
ENSG00000146648	EGFR	Epidermal Growth Factor Receptor

ENSG00000113578	FGF1	Fibroblast Factor 1
ENSG00000105221	AKT2	V-Akt Murine Thymoma Viral Oncogene Homolog 2
ENSG00000133703	KRAS	V-Ki-Ras2 Kirsten Rat Sarcoma Viral Oncogene Homolog
ENSG00000132698	RAB25	RAB25, Member RAS Oncogene Family
ENSG00000087586	AURKA	Aurora Kinase A
ENSG00000145675	PIK3R1	Phosphoinositide-3-Kinase, Regulatory Subunit 1
ENSG00000157554	ERG	V-Ets Erythroblastosis Virus E26 Oncogene Homolog
ENSG00000156802	ATAD2	ATPase Family, AAA Domain Containing 2
ENSG00000134853	PDGFRA	Platelet-Derived Growth Factor Receptor, Alpha Polypeptide
ENSG00000141736	ERBB2	V-Erb-B2 Erythroblastic Leukemia Viral Oncogene Homolog2
ENSG00000214160	ALG3	Alpha-1, 3-Mannosyltransferase
ENSG00000124172	ATP5F1E	ATP synthase F1 subunit epsilon
ENSG00000155097	ATP6V1C1	ATPase, H+ Transporting, Lysosomal 42kDa, V1 Subunit C1
ENSG00000104979	C19orf53	Chromosome 19 Open Reading Frame 53
ENSG00000101266	CSNK2A1	Casein Kinase 2, Alpha 1 Polypeptide
ENSG00000101138	CSTF1	Cathepsin F
ENSG00000136986	DERL1	Derlin 1
ENSG00000185122	HSF1	Heat Shock Transcription Factor 1
ENSG00000125877	ITPA	Inosine Triphosphatase
ENSG00000130312	MRPL34	Mitochondrial Ribosomal Protein L34
ENSG00000114503	NCBP2	Nuclear Cap Binding Protein Subunit 2
ENSG00000186010	NDUFA13	NADH Dehydrogenase (Ubiquinone) 1 Alpha Subcomplex, 13
ENSG00000099795	NDUFB7	NADH Dehydrogenase (Ubiquinone) 1 Beta Subcomplex, 7
ENSG00000147684	NDUFB9	NADH Dehydrogenase (Ubiquinone) 1 Beta Subcomplex, 9
ENSG00000130703	OSBPL2	Oxysterol Binding Protein-Like 2
ENSG00000163882	POLR2H	Polymerase (RNA) II (DNA Directed) Polypeptide H
ENSG00000169398	PTK2	Protein Tyrosine Kinase 2
ENSG00000101146	RAE1	Ribonucleic Acid Export 1
ENSG00000124784	RIOK1	RIO Kinase 1
ENSG00000125870	SNRNP2	Small Nuclear Ribonucleoprotein Polypeptide B2
ENSG00000089006	SNX5	Sorting Nexin 5
ENSG00000271303	SRXN1	Sulfiredoxin 1

ENSG00000104915	STX10	Syntaxin 10
ENSG00000104907	TRMT1	TRNA Methyltransferase 1 Homolog
ENSG00000089195	TRMT6	TRNA Methyltransferase 6 Homolog
ENSG00000185798	WDR53	WD Repeat Domain 53
ENSG00000164924	YWHAZ	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta
ENSG00000147889	CDKN2A	Cyclin-Dependent Kinase Inhibitor 2A
ENSG00000154096	THY1	Thy-1 Cell Surface Antigen
ENSG00000102531	FNDC3A	Fibronectin Type III Domain Containing 3A
ENSG00000136156	ITM2B	Integral Membrane Protein 2B
ENSG00000136143	SUCLA2	Succinate-CoA Ligase ADP-Forming Beta Subunit
ENSG00000138138	ATAD1	ATPase Family, AAA Domain Containing 1
ENSG00000184719	RNLS	Renalase, FAD Dependent Amine Oxidase
ENSG00000165874	FAM35BP	Family With Sequence Similarity 35, Member E, Pseudogene
ENSG00000151702	FLI1	Fli-1 Proto-Oncogene, ETS Transcription Factor
ENSG00000137491	SLCO2B1	Solute Carrier Organic Anion Transporter Family Member 2B1
ENSG00000169554	ZEB2	Zinc Finger E-Box Binding Homeobox 2
ENSG00000163885	CFAP100	Cilia And Flagella Associated Protein 100
ENSG00000181322	NME9	NME/NM23 Family Member 9
ENSG00000162643	WDR63	WD Repeat Domain 63
ENSG00000175826	CTDNBP1	CTD Nuclear Envelope Phosphatase 1
ENSG00000132507	EIF5A	Eukaryotic Translation Initiation Factor 5A2
ENSG00000077150	NFKB2	Nuclear Factor Kappa B Subunit 2
ENSG00000108587	GOSR1	Golgi SNAP Receptor Complex Member 1
ENSG00000131242	RAB11FIP4	RAB11 Family Interacting Protein 4
ENSG00000178691	SUZ12	SUZ12, Polycomb Repressive Complex 2 Subunit
ENSG00000168542	COL3A1	Collagen Type III Alpha 1 Chain
ENSG00000130635	COL5A1	Collagen Type V Alpha 1 Chain
ENSG00000204262	COL5A2	Collagen Type V Alpha 2 Chain
ENSG00000106541	AGR2	Anterior Gradient 2, Protein Disulphide Isomerase Family Member
ENSG00000064270	ATP2C2	ATPase Secretory Pathway Ca2+ Transporting 2
ENSG00000003096	KLHL13	Kelch Like Family Member 13

ENSG00000116138	DNAJC16	DnaJ Heat Shock Protein Family (Hsp40) Member C16
ENSG00000037637	FBXO42	F-Box Protein 42
ENSG00000215695	RSC1A1	Regulator Of Solute Carriers 1
ENSG00000100842	EFS	Embryonal Fyn-Associated Substrate
ENSG00000220785	MTMR9LP	Myotubularin Related Protein 9-Like, Pseudogene
ENSG00000112419	PHACTR2	Phosphatase And Actin Regulator 2
ENSG00000131148	EMC8	ER Membrane Protein Complex Subunit 8
ENSG00000230989	HSBP1	Heat Shock Factor Binding Protein 1
ENSG00000135698	MPHOSPH6	M-Phase Phosphoprotein 6
ENSG00000205356	TECPR1	Tectonin Beta-Propeller Repeat Containing 1
ENSG00000181610	MRPS23	Mitochondrial Ribosomal Protein S23
ENSG00000175662	TOM1L2	Target Of Myb1 Like 2 Membrane Trafficking Protein
ENSG00000198496	NBR2	Neighbor Of BRCA1 LncRNA 2
ENSG00000121152	NCAPH	Non-SMC Condensin I Complex Subunit H
ENSG00000131747	TOP2A	DNA Topoisomerase II Alpha
ENSG00000066279	ASPM	Abnormal Spindle Microtubule Assembly
ENSG00000165480	SKA3	Spindle And Kinetochore Associated Complex Subunit 3
ENSG00000139734	DIAPH3	Diaphanous Related Formin 3
ENSG00000157502	MUM1L1	PWWP Domain Containing 3B
ENSG00000146250	PRSS35	Serine Protease 35
ENSG00000118526	TCF21	Transcription Factor 21
ENSG00000183770	FOXL2	Forkhead Box L2
ENSG00000136574	GATA4	GATA Binding Protein 4
ENSG00000147465	STAR	Steroidogenic Acute Regulatory Protein
ENSG00000090273	NUDC	Nuclear Distribution C, Dynein Complex Regulator
ENSG00000134644	PUM1	Pumilio RNA Binding Family Member 1
ENSG00000065526	SPEN	Spen Family Transcriptional Repressor
ENSG00000129173	E2F8	E2F Transcription Factor 8
ENSG00000165304	MELK	Maternal Embryonic Leucine Zipper Kinase
ENSG00000083290	ULK2	Unc-51 Like Autophagy Activating Kinase 2
ENSG00000154957	ZNF18	Zinc Finger Protein 18
ENSG00000197566	ZNF624	Zinc Finger Protein 624

ENSG00000177138	FAM9B	Family With Sequence Similarity 9 Member B
ENSG00000176746	MAGEB6	MAGE Family Member B6
ENSG00000046774	MAGEC2	MAGE Family Member C2
ENSG00000086289	EPDR1	Ependymin Related 1
ENSG00000219438	FAM19A5	Family With Sequence Similarity 19 Member A5, C-C Motif Chemokine Like
ENSG00000162804	SNED1	Sushi, Nidogen And EGF Like Domains 1
ENSG00000104081	BMF	Bcl2 Modifying Factor
ENSG00000178035	IMPDH2	Inosine Monophosphate Dehydrogenase 2
ENSG00000170500	LONRF2	LON Peptidase N-Terminal Domain And Ring Finger 2
ENSG00000111665	CDCA3	Cell Division Cycle Associated 3
ENSG00000090889	KIF4A	Kinesin Family Member 4A
ENSG00000114796	KLHL24	Kelch Like Family Member 24
ENSG00000171109	MFN1	Mitofusin 1
ENSG00000121864	ZNF639	Zinc Finger Protein 639
ENSG00000141867	BRD4	Bromodomain Containing 4
ENSG00000160877	NACC1	Nucleus Accumbens Associated 1
ENSG00000126003	PLAGL2	PLAG1 Like Zinc Finger 2
ENSG00000121690	DEPDC7	DEP Domain Containing 7
ENSG00000204866	IGFL2	IGF Like Family Member 2
ENSG00000176148	TCP11L1	T-Complex 11 Like 1
ENSG00000126581	BECN1	Beclin 1
ENSG00000173786	CNP	2',3'-Cyclic Nucleotide 3' Phosphodiesterase
ENSG00000141699	RETREG3	Reticulophagy Regulator Family Member 3
ENSG00000006831	ADIPOR2	Adiponectin Receptor 2
ENSG00000073614	KDM5A	Lysine Demethylase 5A
ENSG00000003056	M6PR	Mannose-6-Phosphate Receptor, Cation Dependent
ENSG00000151208	DLG5	Discs Large MAGUK Scaffold Protein 5
ENSG00000135842	FAM129A	Family With Sequence Similarity 129 Member A
ENSG00000163374	YY1AP1	YY1 Associated Protein 1
ENSG00000124782	RREB1	Ras Responsive Element Binding Protein 1
ENSG00000122641	INHBA	Inhibin Subunit Beta A

ENSG00000139329	LUM	Lumican
ENSG00000123815	COQ8B	Coenzyme Q8B
ENSG00000013275	PSMC4	Proteasome 26S Subunit, ATPase 4
ENSG00000160410	SHKBP1	SH3KBP1 Binding Protein 1
ENSG00000111790	FGFR1OP2	FGFR1 Oncogene Partner 2
ENSG00000111731	C2CD5	C2 Calcium Dependent Domain Containing 5
ENSG00000152944	MED21	Mediator Complex Subunit 21
ENSG00000065361	ERBB3	Erb-B2 Receptor Tyrosine Kinase 3
ENSG00000155066	PROM2	Prominin 2
ENSG00000143217	NECTIN4	Nectin Cell Adhesion Molecule 4
ENSG00000088325	TPX2	TPX2, Microtubule Nucleation Factor
ENSG00000175063	UBE2C	Ubiquitin Conjugating Enzyme E2 C
ENSG00000005469	CROT	Carnitine O-Octanoyltransferase
ENSG00000057294	PKP2	Plakophilin 2
ENSG00000091436	MAP3K20	Mitogen-Activated Protein Kinase Kinase Kinase 20
ENSG00000125810	CD93	CD93 Molecule
ENSG00000170989	S1PR1	Sphingosine-1-Phosphate Receptor 1
ENSG00000169744	LDB2	LIM Domain Binding 2
ENSG00000136982	DSCC1	DNA Replication And Sister Chromatid Cohesion 1
ENSG00000172167	MTBP	MDM2 Binding Protein
ENSG00000164754	RAD21	RAD21 Cohesin Complex Component
ENSG00000119865	CNRIP1	Cannabinoid Receptor Interacting Protein 1
ENSG00000140937	CDH11	Cadherin 11
ENSG00000011465	DCN	Decorin
ENSG00000141738	GRB7	Growth Factor Receptor Bound Protein 7
ENSG00000172057	ORMDL3	ORMDL Sphingolipid Biosynthesis Regulator 3
ENSG00000161395	PGAP3	Post-GPI Attachment To Proteins 3
ENSG00000161204	ABCF3	ATP Binding Cassette Subfamily F Member 3
ENSG00000163900	TMEM41A	Transmembrane Protein 41A
ENSG00000101132	PFDN4	Prefoldin Subunit 4
ENSG00000101182	PSMA7	Proteasome Subunit Alpha 7
ENSG00000124164	VAPB	VAMP Associated Protein B And CafY

ENSG00000104375	STK3	Serine/Threonine Kinase 3
ENSG00000104517	UBR5	Ubiquitin Protein Ligase E3 Component N-Recognin 5
ENSG00000179271	GADD45GIP1	GADD45G Interacting Protein 1
ENSG00000254598	CSNK2A3	Casein Kinase 2 Alpha 3
ENSG00000125818	PSMF1	Proteasome Inhibitor Subunit 1
ENSG00000247315	ZCCHC3	Zinc Finger CCHC-Type Containing 3
ENSG00000101193	GID8	GID Complex Subunit 8 Homolog
ENSG00000124217	MOCS3	Molybdenum Cofactor Synthesis 3
ENSG00000189376	C8orf76	Chromosome 8 Open Reading Frame 76
ENSG00000176853	FAM91A1	Family With Sequence Similarity 91 Member A1
ENSG00000170881	RNF139	Ring Finger Protein 139
ENSG00000261236	BOP1	BOP1 Ribosomal Biogenesis Factor
ENSG00000180900	SCRIB	Scribbled Planar Cell Polarity Protein
ENSG00000179886	TIGD5	Tigger Transposable Element Derived 5
ENSG00000101365	IDH3B	Isocitrate Dehydrogenase 3 (NAD(+)) Beta
ENSG00000125901	MRPS26	Mitochondrial Ribosomal Protein S26
ENSG00000215305	VPS16	VPS16, CORVET/HOPS Core Subunit
ENSG00000130311	DDA1	DET1 And DDB1 Associated 1
ENSG00000105393	BABAM1	BRISC And BRCA1 A Complex Member 1
ENSG00000053501	USE1	Unconventional SNARE In The ER 1
ENSG00000197122	SRC	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase
ENSG00000119231	SEN5	SUMO Specific Peptidase 5
ENSG00000105669	COPE	Coatomer Protein Complex Subunit Epsilon
ENSG00000105671	DDX49	DEAD-Box Helicase 49
ENSG00000099797	TECR	Trans-2,3-enoyl-CoA reductase
ENSG00000105135	ILVBL	IlvB Acetolactate Synthase Like
ENSG00000183665	TRMT12	TRNA Methyltransferase 12 Homolog
ENSG00000184402	SS18L1	SS18L1, NBAF Chromatin Remodeling Complex Subunit
ENSG00000149658	YTHDF1	YTH N6-Methyladenosine RNA Binding Protein 1
ENSG00000205981	DNAJC19	DnaJ Heat Shock Protein Family (Hsp40) Member C19
ENSG00000175182	FAM131A	Family With Sequence Similarity 131 Member A
ENSG00000136522	MRPL47	Mitochondrial Ribosomal Protein L47

ENSG00000123908	AGO2	Argonaute 2, RISC Catalytic Component
ENSG00000129292	PHF20L1	PHD Finger Protein 20 Like 1
ENSG00000000419	DPM1	Dolichyl-Phosphate Mannosyltransferase Subunit 1, Catalytic
ENSG00000101166	PRELID3B	PRELI Domain Containing 3B
ENSG00000188428	BLOC1S5	Biogenesis Of Lysosomal Organelles Complex 1 Subunit 5
ENSG00000111845	PAK1IP1	PAK1 Interacting Protein 1
ENSG00000112739	PRPF4B	Pre-mRNA Processing Factor 4B
ENSG00000101247	NDUFAF5	NADH:Ubiquinone Oxidoreductase Complex Assembly Factor 5
ENSG00000132661	NXT1	Nuclear Transport Factor 2 Like Export Factor 1
ENSG00000125846	ZNF133	Zinc Finger Protein 133
ENSG00000125871	MGME1	Mitochondrial Genome Maintenance Exonuclease 1
ENSG00000101343	CRNKL1	Crooked Neck Pre-mRNA Splicing Factor 1
ENSG00000132664	POLR3F	RNA Polymerase III Subunit F
ENSG00000125875	TBC1D20	TBC1 Domain Family Member 20
ENSG00000198356	ASNA1	ArsA Arsenite Transporter, ATP-Binding, Homolog 1 (Bacterial)
ENSG00000179115	FARSA	Phenylalanyl-TRNA Synthetase Subunit Alpha
ENSG00000105607	GCDH	Glutaryl-CoA Dehydrogenase
ENSG00000104957	CCDC130	Coiled-Coil Domain Containing 130
ENSG00000089063	TMEM230	Transmembrane Protein 230
ENSG00000215251	FASTKD5	FAST Kinase Domains 5
ENSG00000125779	PANK2	Pantothenate Kinase 2
ENSG00000120963	ZNF706	Zinc Finger Protein 706
ENSG00000105173	CCNE1	Cyclin E1
ENSG00000147883	CDKN2B	Cyclin Dependent Kinase Inhibitor 2B
ENSG00000106624	AEBP1	AE Binding Protein 1
ENSG00000100234	TIMP3	TIMP Metallopeptidase Inhibitor 3

表2 各基底ベクトルにおける要素値のランクオーダー一覧

