修 士 論 文

マイクロアレイにより得られる遺伝子発現情報からの知識発見に関する研究

指導教官　佐藤賢二 助教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

内藤隆宏

2001 年 3 月

要 旨

目 次

図 目 次

表 目 次

第 1章

はじめに

1953 年に J.Watson と F.Crick によりＤＮＡ構造は２重螺旋モデルあることが提唱されて以来，この半世紀の間に生命科学は猛烈な勢いで進歩した．1980 年代に始まった各種生物の塩基配列解析プロジェクトは，解析技術・機械化・迅速化などの進歩により急速に解析速度が早まった．1995 年に単独で生きる生物として初めてのバクテリアの完全長配列データが，1996 年には初の真核生物として出芽酵母 ($Saccharomyces\ cerevisiae$) の全塩基配列の決定が終了して以来，次々と各種生物の配列決定がなされている．遺伝子も次々に特定され，出芽酵母をはじめ，1 つの生物が持つ全遺伝子セット（ゲノム）が明かなった生物も増えつづけている．現在，全配列が決定されたものは微生物を中心として 37 種類に達している[1] [45]．1988 年から始まったヒトゲノムプロジェクトでは，1999 年に第 22 番染色体の全塩基配列が決定され，それに続き 2000 年に第 21 番染色体の全塩基配列も決定された．さらにドラフト・シーケンス (全塩基配列の約 90%) が公開され，2003 年には完全な全塩基配列が決定される予定であるなど，塩基配列解析の研究に目処が立ってきた．これにより，ゲノム研究の主流はこれまでのシーケンスプロジェクトの成果をもとに，遺伝子の機能の解析を行う，機能ゲノム (Functional Genomics) 研究のパラダイムに突入しようとしている [2].

シーケンスプロジェクトによって得られた成果を解析するとはいっても，対象とする遺伝子の数は数千から数十万にのぼり，さらに，遺伝子の性質をあらわす項目は膨大な数がある．例えば人類にとっても馴染み深い生物であるパン酵母である出芽酵母では，ゲノムサイズは約 1200 万塩基対， 遺伝子数は約 6000 ある [1, 4]．このうち翻訳後の蛋白質の機能が既に明らかになっているもの（機能既知遺伝子）は約 1/3 に過ぎず，全く機能が分からない遺伝子（機能未知遺伝子）も多く含まれる．現在の情報科学的手法による遺伝子

[1] 2000.1.15 現在

の一般的な機能推定は，類似塩基配列がデータベースにあるかどうかホモロジー検索やモチーフ検索などにより調べることである．このため類似する配列が全くなかったり，類似配列があってもそれ自体が機能未知であれば，機能推定を行うことは出来ない [5]．すなわち，データベース内に類似配列がなければ現在の情報科学的手法は全く無力である．また既知の遺伝子との相同性により機能が推定できる遺伝子は，およそ 50 %程度に過ぎないといわれており [6]，配列の類似性のみに注目した機能推定は限界に来ていると思われる．このため，塩基配列情報による機能推定以外の手法を開発することが強く望まれるている．現在，塩基配列情報以外に機能推定に有用と考えられ，利用可能な情報としては遺伝子発現情報，細胞内局在情報，機能上関連, 相互作用情報, 立体構造情報などがある．以前は，出芽酵母細胞のような生物に対して，ゲノム全体にわたって膨大な数の遺伝子の発現情報を得るような解析は不可能であった．しかし，Stanford 大学の P.Brown 研究室によって開発された DNA マイクロアレイを用いることによって [9, 10]，膨大な数の遺伝子について発現の有無を同時に測定することが可能となった．現在では約 6000 個の遺伝子が存在する出芽酵母でさえ，発現解析を行うことが可能となっている [7, 8]．

一方，生物学的な実験技術などの進化により，発現情報以外にも多様な情報が，遺伝子や蛋白質について明らかになってきた．これらの情報を格納したデータベースは一般にゲノムデータベースと総称され，膨大な量のデータが指数関数的に蓄積され続けている．しかし，単にこれらのデータベースから分子レベルの情報を集めて統合化しても，それだけからは遺伝子の機能を推測するのは困難である．このため，ゲノムデータベースを処理する高速なデータ処理機能と知識獲得機能を結合したシステムを構築することの重要性が指摘されている [12, 13, 15, 16]．

しかしながら，マイクロアレイから得られるデータは単にどの遺伝子がどれだけ発現していたかという数値データ列でしかないため，その解析に関して現状ではクラスター分析などが行なわれている程度で，有益な知識を得るための解析手法に関してはまだ手探りの状態といっていい [17]．このような背景により，本研究では，マイクロアレイから得られる遺伝子発現情報にデータベースから得られる機能既知遺伝子の情報を抽出したものを掛け合わせ，ビジネス分野などで利用が行われている相関ルール発見手法を適用することにより [18, 19]，機能未知遺伝子に関する推定をシステマティックに行なうことを試みる．

第 2章

出芽酵母

2.1 はじめに

出芽酵母 (*Saccharomyces cerevisiae*) は，メソポタミア地方では6000 年以上も前からパンの製造に使われ，日本でも清酒の製造に欠かせないものとして使われている．現在でも醸造業・食品業・医療分野等で広く利用されている有用生物である [4]．単細胞の真核生物である酵母は，高等動物の細胞のように発達した細胞内小器官を持ち，遺伝子破壊を始めとした遺伝子操作が容易であることから，現在でも高等動物細胞遺伝子の機能解析研究に広く用いられている [39]．出芽酵母は 16 本の染色体を持ち，それらの合計では 13.5Mb(Mega base:100 万塩基) の長さの塩基配列をもつ．また，配列上に存在する遺伝子の総数は約 6000 である[1] ．表 2.1 に主な生物のゲノムサイズと遺伝子数を示す．

1995 年 7 月に初めて生物の完全長配列データとして，インフルエンザ菌 (*Haemophilus influenzae*) の全塩基配列が決定されたが，その翌年の 1996 年 4 月には 8 年の歳月をかけて行われていた国際プロジェクトによって，出芽酵母のゲノムの全塩基配列が決定された．出芽酵母ゲノム解析研究は，ヨーロッパの研究グループを中心に，米国，カナダ，日本のグループが加わり，全体で 600 人以上の多数の研究者の協力による国際共同研究として行われた [20]．現在では 38 種類の生物の解析が終了し，それ以外にも 77 種類の生物の塩基配列解読が行われている [45]．

[1]同一種であっても，その生物の株が異なればゲノムサイズと遺伝子数は異なる．また年月の経過と共に同定される遺伝子は増えるため，同じ株であっても遺伝子数は異なることがある．

表 2.1: 主要な生物におけるゲノムサイズと遺伝子数 (出典 [44])

生物種	ゲノムサイズ	遺伝子総数
インフルエンザ菌	1.8Mb	1743
マイコプラズマ	0.58Mb	470
枯草菌	4.2Mb	4000
大腸菌	4.7Mb	4000
ラン藻	3.5Mb	2800
出芽酵母	13.5Mb	6000
線虫	100Mb	17800
シロイヌナズナ	100Mb	15000-20000
ショウジョウバエ	180Mb	12000-16000
マウス	3000Mb	80000
ヒト	3000Mb	60000-80000

2.2 モデル生物としての出芽酵母

塩基配列を決定する費用は解析技術の進歩や作業の機械化，解析速度の迅速化などによって下がってはいるが，今なお 1 つの生物のゲノムの全塩基配列を決定するには莫大な予算が必要になる．そのため，闇雲に全ての生物の塩基配列を決定するのでなく，他の生物を研究する上でモデルとなるような有意義な結果が得られることが期待される生物に対象を絞って解析せざるを得ない．このような生物はモデル生物と呼ばれる．そのためゲノム配列決定のプロジェクトでは，分子遺伝学的な研究の成果が蓄積されている微生物や，医学的に価値がある病原微生物，商業的な応用が期待できる有用微生物などのモデル生物を中心に解析が行われている [20].

早期にゲノム解析のモデル生物として出芽酵母が選ばれたのはゲノムサイズの大きさが比較的小さく，有用生物であり，さらに分子遺伝学的な研究の成果が蓄積していたことによる．ゲノム解析以降の酵母は，個々の遺伝子の変異がたやすく作成できるため，変異した遺伝子によっておこる生体の挙動を調べることで，細胞の中で起こることの基本的な解析に広く利用されている．

その他にモデル生物としてよく用いられるものに，線虫，ショウジョウバエ，シロイヌナズナなどがある．これらも出芽酵母と同様に，従来から盛んに実験及び解析が行われ

ている．線虫は，細胞の分化，アポトーシスや初期発生，神経系等の基本的な解析に利用されている．ショウジョウバエの場合は，これらに加えて目や手足がどのようにできるのか，学習や生物時計などのより高次な機能の解析に利用されている．さらに，シロイヌナズナは高等植物の分化や発生の解析に用いられている．

2.3 ゲノム解析により得られたこと

様々な解析が進んでいたと思われていた出芽酵母であるが，塩基配列データの解析によって同定された 6275 個の ORF[2] のうち，約 60%が未知の遺伝子であった．急速に遺伝子の機能解析は進んでいるが，現在でも全く機能が解明されていない機能未知遺伝子が約 1/3 存在している．表 2.2 に出芽酵母の遺伝子の機能分類を示す [20].

表 2.2: 出芽酵母遺伝子の機能による分類 (出典 [20])

機能分類	ORF 数	%
物質代謝関連遺伝子	1145	18.2
エネルギー代謝関連遺伝子	232	3.7
細胞増殖，細胞分裂，DNA 複製関連遺伝子	939	15
転写関連遺伝子	665	10.6
蛋白質合成関連遺伝子	333	5.4
蛋白質修飾，分解などの関連遺伝子	468	7.5
トランスポーター関連遺伝子	314	5
細胞内物質輸送関連遺伝子	356	5.7
細胞構造関連遺伝子	1885	30
情報伝達関連遺伝子	113	1.8
ストレス，細胞障害修復関連遺伝子	296	4.7
トランスポゾンなどモービルエレメント	111	1.8
未知遺伝子	3029	48.3

また，ヒトの疾患との関連が指摘されている遺伝子のうち 30%については，酵母に相同遺伝子が存在していることも明らかにされた．この内のいくつかの相同遺伝子を表 2.3 に示す [44]．今後もヒトゲノムプロジェクトの進展にしたがって，さらに相同遺伝子は増

[2]ORF に関しては次のページを参照

加することが予想されており，医学研究においても，モデル生物としての酵母の重要性はますます高まっていくと考えられる [40].

表 2.3: 疾患関連ヒト遺伝子と酵母遺伝子の類似 (出典 [44])

酵母遺伝子	ヒト遺伝子	酵母における機能	関連するヒト遺伝病
MSH2	MSH2	ＤＮＡ修復蛋白質	遺伝性非ポリポーシス性大腸癌
MLH1	MLH1	ＤＮＡ修復蛋白質	遺伝性非ポリポーシス性大腸癌
YCF1	CFTR	金属耐性蛋白質	のう胞性繊維症
CCC2	WAD	銅イオン輸送蛋白質	Wilson 病（銅蓄積症）
GUT1	GK	グリセロールキナーゼ	グリセロールキナーゼ欠損症
SGS1	BLM	ヘリカーゼ	Bloom 症候群
PAL1	ALD	パーオキシソーム膜蛋白質	Ｘ染色体連鎖性副腎脳白質ジストロフィー
TEL1	ATM	PI3 キナーゼ	毛細血管拡張性運動失調
GEF1	CLCN5	電位作動性塩素チャンネル	Fanconi 症候群
SOD1	SOD1	スーパーオキシド分解酵素	筋萎縮性側策硬化症
YPK1	DM	セリン・スレオニンキナーゼ	筋緊張性ジストロフィー
YIL002C	OCRL	IPP-5 脱リン酸酵素	Lowe 症候群
IRA2	NF1	抑制性調節蛋白質	Ｉ型神経繊維種

2.4 遺伝子発現

生物の生命活動は，遺伝情報を担う DNA と，DNA から作られた蛋白質により成り立っている．この DNA の情報は複製 (replication) され，細胞から細胞へと伝えられる．また細胞内では，DNA 上の遺伝子の部分が蛋白質に翻訳されて，細胞としてのはたらきが行われる．この翻訳は，DNA の情報が直接蛋白質に翻訳されるのではなく，まず DNA の中の遺伝子の部分が RNA に転写 (transcription) される．この RNA には mRNA(messenger RNA), rRNA(ribosomal RNA)，及び tRNA(transfer RNA) の 3 種類がある．rRNA と tRNA は翻訳されて蛋白質にはならないが，細胞の中で様々な機能を果たしている．mRNA のみが翻訳 (translation) されて蛋白質になる．これをとくに遺伝子の発現あるいは遺伝子発現

という [5, 38, 35]．また，蛋白質の生成を担う遺伝子のうち，蛋白質をコードする部分を ORF(open reading frame) とよぶ [37]．この関係を図 2.1 に示す．

図 2.1: 遺伝子発現の流れ

第 3章

マイクロアレイ

3.1 はじめに

各種シーケンスプロジェクトによって次々と全ゲノム配列が決定され，遺伝子も続々と同定されるようになっているが，塩基配列を解析しただけでは十分に遺伝子の機能を解明することは出来ない．そのため，遺伝子の機能を解明する機能ゲノムに研究の焦点が移りつつある．

遺伝子が生体の中で機能するには，図 2.1 で示したように，遺伝子発現という過程を経なくてはならない．逆に言えば遺伝子機能を解明するためには，どのような遺伝子がいつどこで発現しているかが有力な手がかりとなる．この遺伝子発現の解析方法として，従来はノーザンブロット法やディファレンシャルディスプレイ法が用いられていたが，それらの解析遺伝子数はせいぜい 100 前後にすぎない．そのため，シーケンスプロジェクトの成果の有効活用に必要な，ゲノムスケールでの網羅的な遺伝子発現の解析ができなかった [39]．さらに，表 2.1 で示されるようにヒトをはじめとする大規模なゲノムも続々と解読されるようになったため，大量の遺伝子の発現情報を包括的にモニタリングする手法の必要性がますます高まってきた．

このような要求を受けて，Stanford 大学の P.Brown 研究室によって DNA マイクロアレイが開発された [10, 9]．現在，マイクロアレイは急速に医学・薬学・生物学などの研究分野で普及し，ポストゲノム研究の中でも最も注目される技術の 1 つとして大きな成果が期待されている [39]．

3.2 マイクロアレイの原理

図2.1に示したように，遺伝子の発現は，DNAの情報がmRNAに転写される段階(mRNA合成)と，転写されたmRNAの情報が蛋白質に翻訳される段階(蛋白質合成)に分けられる．マイクロアレイはこのmRNAの転写量の変化を数万個の遺伝子について同時に観測可能にする実験手法である．

マイクロアレイ技術には，主に2つのタイプがある．1つはオリゴヌレクチオドによるGene Chipで，Affymetirx社が開発販売を行っており，オリゴDNAをガラス表面上で合成していくタイプである．もう1つはcDNAを貼り付けていくタイプのものでDNAマイクロアレイと呼ばれるものである．これは，1995年にStanford大学のP.Brown研究室によって，シロイヌナズナの遺伝子発現解析への応用が最初に報告された[10]．Gene Chipは半導体製造に用いられる光リソグラフィ技術により作成されるため，コンピュータのマイクロチップ製造工程との類似からDNA Chipと呼ばれることもある．両者ともに蛍光色素で標識した核酸を付加し，これがどの程度発現するかを観測することにより大量の遺伝子の発現変化を解析することが出来る[42]．以下本論文では，マイクロアレイはP.Brown研究室のマイクロアレイを指すことにする．表3.2にGene ChipとDNAマイクロアレイを比較したものを示す[11]．

表 3.1: GeneChipとDNAマイクロアレイの比較

	Gene Chip	DNAマイクロアレイ
開発元	Affymetirx社	Stnaford大学P.Brown研究室
作成法	基板上でDNAを合成	基盤上にDNAを固定
技術	光リソグラフィ	DNAスポッター, スライドガラス
DNA	一本鎖オリゴDNA	通常cDNA
区画形状	正方形($20\mu m \times 20\mu m$)	円形(直径100 – 400μ m)
集積度	–300,000/チップ(現状)	–10,000/アレイ(現状)
応用領域	遺伝子の変異・多型性の検出 遺伝子発現モニタリング	遺伝子発現モニタリング

3.3 データ処理

現在の一般的なマイクロアレイの実験では定量的にmRNAの転写量を測定することが出来ない．これは，実験に使用する蛍光物質の違い，スライドガラスのロットの差，試料の調整の差の違いによって観測するべき蛍光強度が本来のものと異なってしまうためである [49]．そこで野生株 (wild) と対象となる遺伝子破壊株 (mutant) を同一条件下で実験を行い，発現量の比をとることで相対的な比較を行う．この発現比の分布は右に裾の長い非対称分布であり，対数正規分布として近似できる．そこで本研究では発現量の比を対数によって正規化することを考える．式3.1に対数正規分布の式を示す．マイクロアレイから得られたデータに対数正規分布確率密度関数を当てはめた例を図3.1に示す．この対数によるデータの正規化はマイクロアレイの解析に広く用いられている [50].

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp(\frac{-(\log x - \mu)^2}{2\sigma^2}) & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{3.1}$$

図 3.1: マイクロアレイと対数正規分布確率密度関数

また，ゲノムスケールで網羅的に解析を行う場合，wild と mutant のどちらかもしくは両方の発現量が低い場合，比を取ったときに極端な値が出てしまうおそれがある．本研究では，ユーザが次に示す３つの信頼限界のパラメーターを設定することにより，異常なデータを排除することを考える．

1. wild の最小値

2. mutant の最小値

3. mutant/wild の最小値

本研究で用いているデータの一例を表3.2 に示す．このデータは，九州大学大学院の久原研究室から提供されたものである．

表 3.2: マイクロアレイのデータ

ORF	mutant	wild	mutant/wild	$\cdots$
YAL003W	2168.553	1693.21143248892	0.780802421010195	$\cdots$
YBL078C	68.062	31.9124032209149	0.468872545927462	$\cdots$
YCR105W	0.081	16.902036543571	208.667117821864	$\cdots$
YAR068W	221.632	95.2922957516728	0.429957297464594	$\cdots$
YBL107C	81.127	101.78978422113	1.25469676212765	$\cdots$
YHL002W	1.118	27.8553509697438	24.9153407600571	$\cdots$
YDR134C	3932.009	5887.47296694947	1.49731930088397	$\cdots$
YDR139C	578.38	274.341052452529	0.474326657997388	$\cdots$
YDR252W	67.702	50.7880762603587	0.75017098845468	$\cdots$
YCL055W	67.809	204.832219965625	3.02072320732683	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

3.4 解析手法

3.4.1 はじめに

マイクロアレイは従来の方法に比較して，大量の遺伝子の発現情報を得る事が出来る．しかし網羅的に全てを解析しようとすればそのデータ量はさらに大きなものになる．特に

図??のような，遺伝子発現情報の時間変化であるタイムコースを得ようとすればデータ量は膨大なものになる．たとえば，酵母の全遺伝子を対象として，100 時点のデータを解析したとすれば $6,000(\text{全遺伝子}) \times 100(\text{パターン}) = 600,000$ となる．そのためマイクロアレイデータから情報科学的な手法によってシステマティックに解析することが望まれており，現在クラスタリングを中心に開発が行われている．

図 3.2: タイムコースの例

3.4.2 クラスタリング

現在，マイクロアレイの解析においてもっとも広く行われている手法の 1 つがクラスタリングである [52]．クラスタリングとは，タイムコースが似通った遺伝子は，細胞内でも似通った機能を担っているという生物学的背景に基づいた推測手法である．実際，出芽酵母の既知遺伝子 2467 個について同様の機能を担う遺伝子は同一のクラスターに分類されることが示されている [53]．そのためマイクロアレイの実験において共通して発現している一群の遺伝子群を同定し，既知の遺伝子の生物学的な注釈を頼りにして，同じクラスターに入っている機能未知遺伝子の機能を予想するというのが一般的な手法である．図 3.2 を発現・非発現の 2 値表現を行ったものを表 3.3 に示す．これは，発現パターンと呼

ばれる．たとえば，臓器毎の挙動に関して複数の遺伝子の発現量を時間軸に沿って観測する．発現パターンの類似性に基づいて遺伝子をクラスタリングによってグループに分類する．こうすることで，分類されたグループ内の機能既知遺伝子の性質から，機能未知遺伝子の性質を類推できると考えられている．このような考え方に基づき，現在，いくつかの解析手法の開発が行われている [47, 17, 41].

表 3.3: 発現パターンの例 (発現状態を 1，非発現状態を 0 で表す．)

Time(Sec)	5	10	15	20	30	40	60
gene1	0	0	1	1	0	0	1
gene2	0	1	1	1	1	1	1
gene3	1	1	1	1	1	0	1
gene4	1	1	1	1	1	1	1
gene5	0	1	1	1	0	0	1
gene6	0	0	0	0	1	0	0
gene7	0	0	1	1	1	1	1

3.4.3 問題点と改善点

機能既知遺伝子との発現パターンの類似性をもとに行う遺伝子機能予測は，塩基配列の相同性から機能類推が出来ない場合，特に有効な解析手法になりうると考えられる．しかし，既存の多くの解析手法では未知遺伝子の機能を推測するのに参照すべき発現パターンが類似する既知遺伝子の機能注釈付け (Functional Annotation) 情報がないばかりに有効な推測が出来なかった．そこで本研究では複数のデータベースから多くの機能注釈付けを得ることによって，より有効な解析を得ることができると考える．

第 4章

データマイニング

4.1 データマイニングが登場した背景

近年，巨大なデータベースから知識獲得を高速に行う技術としてデータマイニングがデータベースと人工知能の境界分野で注目されている．データマイニング (data mining) という言葉は，巨大なデータベースを鉱山にたとえ，その中から何らかの法則 (知識) を宝として発掘するという意味でつくられた．データマイニングが登場した背景として技術革新によって情報処理装置，情報記録装置及び情報周辺機器が広く普及し，POS などによって大量の電子データが迅速に収集されるようになったことが挙げられる．実際に，売上情報，顧客情報，地図情報，生物情報などの巨大データベースが存在する．これらの巨大なデータに対して，非定型的な検索や集計，可視化などの伝統的な処理を高速に実行したいという要求に応えようとしているのが，データウェアハウスや OLAP[1] などである．さらに進んで，マーケティングや危険度予測などに役立つ知識を抽出したいという要求に応えようとしているのが，データマイニングである [28].

データマイニングシステムは，例えばスーパーマーケットのレシート情報から,「商品 A と B を購入した顧客は高い確率で商品 C を購入する」という形をした法則 (ルール) を収集するものである．すると，ある商品 C を重点的に売るためには他のどのような商品 A と B との組み合わせが有効であるとか，商品 A，B を販売しなくなると商品 C にどのような影響が出るかを網羅的に理解できる．このようなルールは関係データベースの問い合わせ文を大量に使っても生成出来るが，巨大なデータベースに対してこの操作を行うと，膨大な時間を消費してしまう．そのため，次の 2 つの課題を満たすデータマイニングアル

[1] オンライン分析処理 (On-Line Analytical Processing)．多次元データベースや多次元データ分析の意味で用いられる．

ゴリズムが必要になる．1 つは，巨大なデータベースの情報を効率的に扱うことが出来るアルゴリズムであること．もう 1 つは，膨大なデータトランザクションの属性間やトランザクション間に存在する関係から，網羅的に全てのパターンや規則性を見つけ出し，人間が理解できる形式で結果を示すことができるアルゴリズムでなくてはならないということである．

従来から用いられている決定木やクラスター分析は，トランザクション数にして高々数万程度の比較的小規模なデータベースに対して主に適用されており，システムとしても数千万レコードからなる巨大なデータベースに適用することを想定して作られてはいない [28]．一方，相関ルール発見は網羅的なルールを高速に生成するアルゴリズムとして 1994 年に IBM のアルマデン研究所で開発された．これにより，10 万トランザクション程度の比較的小さなデータベースなら，数十分の 1 から数百分の 1 の計算時間でルールが生成出来るようになった．したがって，大規模なデータを扱える相関ルール発見は，データマイニングの中でも極めて有効な方法であると考えられる．

生命科学分野では，最近の急激な発展により大量の生命情報がデータベースに蓄積されるようになってきている．日本では，1991 年より開始されたヒトゲノムプロジェクトにより，京都大学化学研究所と東京大学医科学研究所ヒトゲノム解析センターが中心となって，ゲノムネット (GenomeNet) と名づけられたコンピュータネットワークの構築，整備，運用を行ってきた．これは，生物学分野の文献情報やゲノム地図，塩基配列，蛋白質のアミノ酸配列及び立体構造，代謝系や制御系の分子ネットワーク，神経系や免疫系における細胞のネットワーク，発生・分化・疾病に関するデータなど多種多様なデータを蓄積している生命科学情報の統合データベースである図 4.1 で示すようにゲノムネットに蓄積されている個々のデータベースの情報量は爆発的な勢いで増加しており，従来の方法では十分な解析が出来なくっている．そのため，大量のデータの中から法則を見つけ出せるデータマイニングが大きく注目されている．特にマイクロアレイのような大量の遺伝子発現データを扱うデータマイニングは”gene mining”と呼ばれ，活発に研究が行われている [56]．

4.2 データマイニングのプロセス

データマイニングのプロセスは，図 4.2 のように，7 つの段階からなる．データマイニングの手順は左側から入り，右に順々に移動し，そして報告の段階に進んでいく．

下に図 4.2 の各段階を説明する．

- **Step.1 目標の決定**

  ユーザが得たいと考える課題を明確に定義する．

図 4.1: 主要ゲノムデータベースのデータ量の遷移

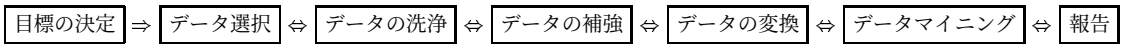

図 4.2: データマイニングのプロセス

- **Step.2 データの選択**
  与えられた生データから，どのような項目を使うか選択する．
- **Step.3 データの洗浄**
  選択したデータの品質を確保する為に欠損値と異常値の排除あるいは補完を行う．
- **Step.4 データの補強**
  与えられた生データの他に，他の使用可能なデータベースにより生データの情報を補強する．
- **Step.5 データの変換 (コード化)**
  データをデータマイニング用に適した形式に変換する．これから行う分析，およびデータマイニングアルゴリズムに必要なデータ形式にあわせて，データを変換する．
- **Step.6 データマイニング**
  変換されたデータを用いてマイニングを行う．これはプロセスの中核を担う段階であり，適切なデータマイニング手法を組み合わせを選択することを除けば，一般にこの処理は高速である．

- **Step.7 報告**
  データマイニングによって得られた結果をユーザに提示する.

一貫した流れがあるように思われるが，実際には全ての段階で，1 つあるいはもっと前の段階に戻ることがある．これは，データマイニングの段階で，洗浄が不十分であったと気づいてやり直したり，データの補強が十分ではなく新たなデータを加える必要がある場合などがあるからだ.

データマイニングを使って実際に知識発見が行われるのは Step.6 であるが,「ゴミを入れればゴミが出る」といわれように，データマイニングはデータの量と質に大きく依存する．どんなに優れたデータマイニング手法であっても，良質のデータがなければ有益な結果は得られない．そのため，データの選択，データの洗浄およびデータの補強などのデータ前処理段階こそが，データマイニングを行う，うえで非常に重要なプロセスになる．実際，表 4.1 に示すように多くの作業がデータマイニングの前処理に費やされることからも，この作業の重要性が示されている.

表 4.1: データマイニングプロセスに必要な作業量 (出典 [25])

プロセス	作業量
目標の決定	20%
データ準備	50%
データマイニング	10%
結果の分析と理解	10%

また，理想的な状況でさえ，データマイニングのプロセスは，完成することのないプロセスであるといわれる．データマイニングのプロセスにおける全ての段階で，あらゆる可能性を追求し工夫をこらすことにより，データマイニングのプロセスはより洗練されたものになる [23].

4.3 データマイニングの手法

データマイニングは，統計手法，機械学習，可視化技術，データベース技術など様々な分野の融合分野であり，種々のアルゴリズムがある．その内のいくつかの手法を以下に示す [19].

- クラシフィケーション (Classification)–特徴付け
  データをあらかじめ定められたいくつかのクラスに分類する関数を学び，既知データから規則性を発見し，それをもとに未知データの判別，予測などを行う手法である．分類の意味づけは，この手法を利用する．決定木やニューラルネットなどが代表的な手法である．

- クラスタリング (Clustering)–分類
  データを記述する有限個のカテゴリーの集合を同定するものであり，分類基準が分かっていないデータを，データの類似性からいくつかのグループに自動的に分類する手法である．クラスタリングでは，グループの分け方をあらかじめ設定しなくてもよいという点がクラシフィケーションとは異なる．統計分析や教師なし学習などが主要な手法である．代表的な手法として k-Nearest Neighbor などがある．

- アソシエーション (Association)–関連性抽出
  データ属性の間に存在する相関関係を学び，関連性の強いデータの組合せパターンを検索する手法である．相関ルール発見などが代表的な手法である．

これらの中から以下の観点に基づきデータ性質や結果に応じて，適切な手法を選択する [24].

- 適用の一般性
  どれだけ多くの問題解決やデータ型に適用することができるか.

- 結果の透過性
  データマイニングによって得られた結果が判りやすく，理解しやすいことは重要である．このような結果は透過性があると言われる．マイニングによって得られる結果が透過性を持つものと，そうではないものがある．例えば，決定木や相関ルール発見は明確なルールをもたらすため，わかりやすく透過性のある結果が得られる．しかし，ニューラルネットワークやクラスタリング手法では特定のモデルがなぜ得られたかについてほとんど教えてくれないため，透過性のない結果が得られる．

本研究では結果の透過性が優れており適用の一般性もある，相関ルール発見手法を用いた．相関ルール発見は基本的に属性の積集合演算を行うので，種類や性質の異なる複数のデータソースを用意に統合し，マイニングすることが可能である．さらに分類属性を明確に定めなくてもよいので，分類属性を定めることが出来ないものを対象として，網羅的にマイニングすることが可能である．

4.4 相関ルール発見

4.4.1 はじめに

相関ルール発見は，もともとビジネス分野への応用を意識して開発された手法である．例えば，小売業ではPOSデータを，通信販売業では顧客の過去の購買記録などをデータベース化して蓄えている．そこで，購買記録から求め，どのような商品 (アイテム) を組み合わせたトランザクションが多いかを計算し，相関ルールを求めることにより商品の配置やカタログの構成などにおいて望ましい組み合わせを発見する．このように，購買傾向に対して成立する相関ルールを用いれば，マーケティングを行う上で有効である．このような理由でダイレクトメール送付に関する戦略決定など多くの分野で利用されている．図 4.3 に相関ルール発見の例を示す.

	ミネラルウォーター	牛乳	パン	おにぎり
客1	買う	買わない	買う	買う
客2	買わない	買う	買わない	買わない
客3	買う	買う	買う	買わない
客4	買う	買わない	買う	買わない
客5	買わない	買わない	買う	買う

$\longrightarrow$ ミネラルウォーターを買う人はパンを買う人が多い

図 4.3: 相関ルールの例

4.4.2 相関ルールの定義

アイテム集合を$\tau = \{i_1, i_2, i_3, \cdots, i_m\}$, トランザクションデータベースを$\mathcal{D} = \{t_1, t_2, t_3, \cdots, t_n\}$, $t_i \subseteq \tau$ とする． 各要素をt_iをアイテム集合 (itemset) と呼ぶ．長さkのアイテム集合とはkのアイテムの組み合わせを指す.

相関ルールは$Y \Leftarrow X$で表現される．ただし，$X, Y \subset \tau$ かつ$X \cap Y = \emptyset$とする．ルール$Y \Leftarrow X$の右辺であるXをルールの本体 ($body$)，左辺Yをルールの頭部 ($head$) と呼ぶ．相関ルールは支持度 (support)，確信度 (confidence) の2つのパラメータを有し，これらの値によって相関ルールの価値が表現される．相関ルールは常に2値属性上で定義される[2] ．相関ルール$Y \Leftarrow X$の支持度$support(Y \Leftarrow X)$は$\mathcal{D}$全体に対しXとYをともに含むトラン

[2]現在では連続値が扱えるように拡張されているが，本論文では2値の場合だけを考える.

ザクションの割合 $support(Y \cup X)$ により定義される．また，確信度 $confidence(Y \Leftarrow X)$ は $\mathcal{D}$ の中で X を含むトランザクションのうち，X と Y を共に含むトランザクションの割合，すなわち $support(Y \cup X)/support(X)$ によって定義される．書き換えると，式 4.1，式 4.2 のようになる．

$$\text{支持度} = \frac{\text{アイテム集合 } X \text{ と } Y \text{ を共に含むトランザクション数}}{\text{全トランザクション数}} \tag{4.1}$$

$$\text{確信度} = \frac{\text{アイテム集合 } X \text{ と } Y \text{ を共に含むトランザクション数}}{\text{アイテム集合 } X \text{ を含むトランザクション数}} \tag{4.2}$$

相関ルールの抽出問題はユーザによって指定された**最小支持度** (minimum support) と**最小確信度** (minimum confidence) を満足する全てのルールを見出すことに相当する．

4.4.3 抽出アルゴリズム

相関ルールは次の 2 つのステップで抽出される．

- **Step.1** 最小支持度を満足するアイテム集合をすべて見つけ出す．これらのアイテム集合をラージアイテム集合と呼ぶ．
- **Step.2** Step.1 で求めたラージアイテム集合から，最小確信度を満たす相関ルールを導き出す．

相関ルール抽出処理のうち，最小支持度による足きりの問題を除けば Step.1 は可能な全てのアイテム集合について支持度を計算する処理であるといえる．トランザクションデータベースを繰り返しスキャンし，アイテム集合の支持度を計算するため，アイテム数が多くなると組み合わせ論的にアイテム集合のバリエーションが増え，それに伴い高い計算負荷がかかる．一方，Step.2 は Step.1 で最小支持度を超えたラージアイテム集合だけを対象に相関ルール発見を行うため，Step.1 と比べるとその負荷は軽い．そのため，非常に低い支持度および確信度を持つルールまで探そうとすると，その数は爆発する．従ってユーザはまず適切な最小支持度および最小確信度を与えることが求められる．

4.4.4 アプリオリ

アプリオリは相関ルール発見に現在最も広く利用されているアルゴリズムで，1994 年に IBM アルマデン研究所の R.Agrawal らによって開発された [18]．これを基にその後，多

くのアルゴリズムが研究されてきた．図 4.4 にアルゴリズムを示す．

ここで，k 個のアイテムの組み合わせを k-itemset，長さ k のラージアイテム集合を L_k，長さ k の候補アイテム集合を C_k とする．長さ $k(\geq 2)$ の場合の処理は次のようになる．

- **Step.1** 長さ $(k-1)$ のラージアイテム集合 L_{k-1} から，長さ k の候補アイテム集合 C_k を作成する．
- **Step.2** トランザクションデータベースを検索し，支持度を求める．
- **Step.3** 最小支持度を満足するものを取り出し，長さ k のラージアイテム集合 L_k とする．

上記の長さ k のラージアイテム集合を求める処理をパス k と呼ぶ．この処理は新たなラージアイテム集合が空となるまで続けれられる．ここで，長さ $(k-1)$ のラージアイテム集合から長さ k の候補アイテム集合を生成する手順を図 4.5 に示す．**Join Step** は SQL の結合演算を用いて簡潔に表現している．その後，**Prune Step** で長さ $(k-1)$ のラージアイテム集合に含まれない組み合わせを持つものを捨て去り，残りを長さ k の候補アイテム集合にする．

Algorithm 1. *apriori*

$L_1 \leftarrow$ large 1-itemsets
$k \leftarrow 2$
while$(L_{k-1} \neq 0)$ **do begin**
    $C_k \leftarrow L_{k-1}$ から生成された候補アイテム集合
    forall データベース $\mathcal{D}$ のトランザクション t **do begin**
        t に包含される C_k 中のすべての候補の計数を 1 増加する
    end
    $L_k \leftarrow C_k$ 中の最小支持度を満たす候補
    $k \leftarrow k+1$
end Answer $\leftarrow \bigcup_k L_k$

end

図 4.4: アプリオリアルゴリズム

1. **Join Step**

   insert into candidate k-itemsets

   select p.item$_1$, p.item$_2$,$cdots$, p.item$_{k-1}$, q.item$_{k-1}$

   from large $(k-1)$-itemset, p, large$(k-1)$-itemset q

   where p.item$_1$=q.item$_1$, $cdots$, p.item$_{k-2}$=q.item$_{k-2}$, p.item$_{k-1} < q$.item$_{k-1}$;

2. **Prune Step**

   forall itemsets $c \in$ candidate k-itemset **do**

   forall (k-1)-subsets s of c **do**

   if ($s \notin$ large(k-1)-itemsets) **then**

   delete c from candidate k-itemset;

図 4.5: アプリオリアルゴリズムにおける集合の導出

例えば，マイクロアレイの実験によって表 4.2 のようなデータが得られたとする．この表は以下のことを表している．

- gene1 は実験 2 で発現し，機能 a と b を有する．
- gene2 は実験 1 で発現し機能 a と b と c を有する．
- gene3 は実験 2 で発現し，機能 c を有する．
- gene4 は実験 1 と 2 で発現し，機能 a と c を有する．

このままでは扱いづらいので表 4.3 のように置き換えて閾値は，最小支持度を 50%，最小確信度 75%とするして計算すると，表 4.4 が導出される．この実行例は表 4.5 にまとめて示す．

表 4.2: 仮想的なデータ

トランザクション ID	アイテム
gene1	exp2, function1, function3
gene2	exp1, function1, function2, function3
gene3	exp2, function3
gene4	exp1, exp2, function1, function3

表 4.3: トランザクションデータベース

トランザクション ID	アイテム
1	b, c, e
2	a, c, d, e
3	b, e
4	a, b, c, e

表 4.4: ラージアイテム集合 (最小支持度 50%)

ラージアイテム集合	支持度
a	75%
b	50%
c	50%
a, c	50%

表 4.5: アプリオリによるラージアイテム集合の例

Databse $\mathcal{D}$

TID	Items
100	a, c, d, e
200	b, c, e
300	a, b, c, e
400	b, e

Scan $\mathcal{D}$ $\longrightarrow$

C_1

Itemset	Count
{a}	2
{b}	3
{c}	3
{d}	1
{e}	3

L_1

Itemset	Count
{a}	2
{b}	3
{c}	3
{e}	3

C_2

Itemset
{a, b}
{a, c}
{a, e}
{b, c}
{b, e}
{c, e}

Scan $\mathcal{D}$ $\longrightarrow$

C_2

Itemset	Count
{a, b}	1
{a, c}	2
{a, e}	1
{b, c}	2
{b, e}	3
{c, e}	2

L_2

Itemset	Count
{a, c}	2
{b, c}	2
{b, e}	3
{c, e}	2

C_3

Itemset
{b, c, e}

Scan $\mathcal{D}$ $\longrightarrow$

C_3

Itemset	Count
{b, c, e}	2

L_3

Itemset	Count
{b, c, e}	2

- **Step.1** $minisupp$ を計算する．
  与えられたデータのトランザクション数が 4，最小支持度が 50% であるから，$minisupp = 2$ となる．

- **Step.2** Database $\mathcal{D}$ より C_1 を生成する．
  トランザクションデータベースを検索し，それぞれのアイテムがトランザクションに含まれる回数を数え上げ，その結果を C_1 に格納する．

- **Step.4** C_1 から L_1 を生成する．
  C_1 のアイテム集合 {{a}, {b}, {c}, {d}, {e}} から，Step.1 で求めた $minisupp = 2$ を満たすもののみ {{a}, {b}, {c}, {e}} を取り出し，これをラージアイテム集合として L_1 に格納する．

- **Step.5** L_1 から C_2 を生成する．
  ラージアイテム集合 L_1 ={{a}, {b}, {c}, {e}} から 2 つのアイテムの組み合わせ，{{a, b}, {a, c}, {a, e}, {b, c}, {b, e}, {c, e}} を生成し，候補アイテム集合として C_2 に格納する．

- **Step.6** C_2 から L_2 を生成する．
  候補アイテム集合 C_2 の支持回数 (それぞれのアイテムがトランザクションに含まれる回数) をデータベースを検索して数え上げ，長さ 2 のラージアイテム集合 {{a, c}, {b, c}, {b, e}, {c, e}} を求め L_2 に格納する．

- **Step.7** L_2 から C_3 を生成する．
  ラージアイテム集合 L_2 ={{a, c}, {b, c}, {b, e}, {c, e}} から 3 つのアイテムの組み合わせ {b, c, e} を生成し，候補アイテム集合として C_3 に格納する．

- **Step.8** C_3 から L_3 を生成する．
  候補アイテム集合 C_3 の支持回数をデータベースを検索して数え上げ，長さ 2 のラージアイテム集合 {b, c, e} を求め L_3 に格納する．

- **Step.9** 同様の手順で L_3 ={b, c, e} から C_4 を生成しようとしても 4 以上のメンバーが存在しないので終了となる．

以上により，L_1，L_2，L_3 が最小支持度を満たすラージアイテム集合となる．

次に，これらのアイテム集合の確信度を求める．最小支持度を満たしているラージアイテム集合は，L_1，L_2，L_3 である．表 4.5 から，ラージアイテム集合は，

- L_1={{a}, {b}, {c}, {e}}
- L_2={{a, c}, {b, c}, {b, e}, {c, e}}
- L_3={b, c, e}

と書き換えることができる．最小支持度と最小確信度を満たしているものとして，$c \Leftarrow a$，$e \Leftarrow b$，$e \Leftarrow b, c$ など合計 16 のルールが導出される．これらの例をもとに先ほどの表 4.2 に当てはめ直してみると，

- 実験 1 で発現している遺伝子は機能 a を有する．
- 実験 2 で発現している遺伝子は機能 b を有する．
- 実験 2 で発現し機能 a を有する遺伝子は機能 c を有する．

という相関ルールが得られたことになる．

4.5 相関ルール発見の問題点

相関ルール発見の，問題点はいくつかあるが，ここでは計算量，ルールの透過性，ルールのふるい分けに関して述べる．

4.5.1 計算量

典型的なコンビニエンスストアでは 2800 のアイテムを扱っている [29]．この場合，2 つのアイテムの組み合わせだけも 3918600 になり，3 つのアイテム組み合わせでは 3654747600 の組み合わせとなり，組み合わせの爆発が起こってしまう．

現在の計算機の能力には以前と比べて劇的に向上しているが，これだけの組み合わせを計算するには依然として，計算能力が不足している．本研究でも同様のことがいえる．本研究ではマイニングデータの前処理段階を工夫することによって，この組み合わせ爆発をある程度抑えることが出来た．これに関しては，第 5.3.1 章で述べる．

4.5.2 ルールの透過性

相関ルール発見では，最小支持度と最小確信度の 2 つのパラメータによって抽出されるルールを絞り込むが，抽出されたルールの中にはしばしば無意味なルールが含まれている．次の 3 つの事例を考える．

1. 木曜日にスーパーマーケットでビールと紙おむつを一緒に買う人が多い.

2. 顧客は製品保証をつけた大型の家電製品を買う傾向がある.

3. 日曜大工店の新規オープンでよく売れるものの 1 つがトイレットリングである.

この 3 つの事例は，それぞれ相関ルール発見でもたらされる「有益なルール」,「自明なルール」,「説明不可能なルール」の 3 つを表している.

- 有益なルール
  有益なルールは品質の高い実行可能な情報である．1. の事例は木曜日の晩に子供のおむつと父親のビールを週末用に準備することを示している．店側がビールのある通路の近くにおむつを置いておくと，2 つの商品はさらに売上を伸ばすことがことができるのである．このルールから発展して次のような戦略を立てることも可能である．例えば，顧客が買い忘れないようにビールの見える範囲に他のベビー向け商品を置くとか，他のレジャー関連の食品をベビー用品売り場に置く等が考えられる.

- 自明なルール
  自明なルールは，その業界の人なら誰でもすでに知っているルールのことである．我々は，すでに大型家電製品の購買と同時に保証契約をつけることを知っている．保証契約は大型家電製品と一緒に広告されており，単独で使われることはめったにない．このようなルールは，データ上は正しいが，使いようがないのである．同様な結果はたくさんある．ペンキを買う人はペンキ用のブラシを買うあるいはシャンプーとリンスは同時に買われるなどと同じである.

- 説明不可能なルール
  説明不可能なルールとは不可解な結果であり理解することが出来ず，それに対して対応することが難しいものである．3. の事例は，新事実の発見ではないかと期待させるものであるが，消費者行動や商品についての洞察がかけており，次の行動を示唆することは出来ない．開店セールの間，トイレットリングが他の商品よりも安かった，あるいは少数店舗だけの例外的な事例なのかなどいろいろと考えられる．このようなルールは，原因がどのようなものであれ，相関ルール発見に用いたデータからの追加分析でも確実に説明することは出来ない.

相関ルール発見を行うときには，多くの結果が自明なルールであるか，説明できないルールであることが多い．どのルールが価値あるものかを知るためには，事前のデータに対する背景知識を有していることが必要とされる [24].

4.5.3 ルールの精錬

巨大なデータベースから抽出される相関ルールはしばしば大量で，かつ各々のルールは高々数個のアイテムの相関を示すだけであり，全体を一度に把握することが困難である．発見される相関ルールの中には，他の相関ルールから自然に導き出されてしまうような，冗長なルールともいえるルールが多く含まれる場合がある．このような「面白くない」ルールは統計的検定を用いて除去することが出来る．本研究では，ルールのボディとヘッドの独立性に注目したルールの精錬手法を用いている．以下に，その方法を説明する．

例えば,「ミネラルウォーター $\Leftarrow$ 牛乳」というルールが存在し，次のように支持度がなっていたとする．

$$\begin{aligned} support(\text{牛乳}) &= 16\% \\ support(\text{ミネラルウォーター}) &= 25\% \\ support(\{\text{牛乳}, \text{ミネラルウォーター}\}) &= 4\% \end{aligned}$$

この例では，牛乳の支持度にミネラルウォータの支持度を掛けた値が { 牛乳, ミネラルウォーター } のサポートに等しくなっており $(16\% \times 25\%)$，牛乳とミネラルウォータは独立している．このルールはヘッドとボディに正の相関があるルールではないので「面白くない」ルールである．そこで，一般にルール $H \Leftarrow B$ に対して統計的検定を用いて「B と H が独立である」という仮説が棄却できれば有意なルールとし，そうでなければ「面白くない」ルールとすることができる．この検定のためにまず，トランザクションの総数を N，アイテム集合 X の支持度 $support(X)$ として，表 4.6 のような表を作成する．ここで，観測度数とは条件を満たすトランザクションが実際に発生した回数を表し，期待度数とは H と B が独立事象であると仮定したときに条件を満たすトランザクションの発生が予想される回数である．このとき式 4.3 は，自由度 1 の $\mathcal{X}^2$ 分布に従うことが知られている．もし T_{dep} が 0 に近ければ H と B は互いに独立であり，大きければ H と B は互いに相関が強い．ユーザが与えた有効水準 α を元に ${T_{dep}} < {\mathcal{X}_1}^2(\alpha)$ であれば H と B が独立であるとみなして，ルール $H \Leftarrow B$ は「面白くない」ルールであるとする．例えば，有意水準 α を 5%とすると，${T_{dep}} < {\mathcal{X}_1}^2(0.05) = 3.841$ であればそのルールは「面白くない」ルールとして捨てる [26].

$$\begin{aligned} T_{dep} &= \sum \frac{(\text{観測度数} - \text{期待度数})^2}{\text{期待度数}} \\ &= N \frac{(support(\{H, B\}) - support(H)support(B))^2}{support(H)support(B)(1 - support(H))(1 - support(B))} \end{aligned} \tag{4.3}$$

表 4.6: 独立検定用の分割表

条件	観測度数
	期待度数
$B \wedge H$	$Nsupport(\{H, B\})$
	$Nsupport(H)support(B)$
$B \wedge \neg H$	$N(support(B) - support(\{H, B\}))$
	$Nsupport(B)(1 - support(H))$
$\neg B \wedge H$	$N(support(H) - support(\{H, B\}))$
	$Nsupport(H)(1 - support(B))$
$\neg B \wedge \neg H$	$N(1 - support(B) - support(H) + support(\{H, B\}))$
	$N(1 - support(B))(1 - support(H))$

4.6 決定木

相関ルール発見は離散値しか扱えないのでマイクロアレイから得られる連続値の遺伝子発現状態を「正に発現」,「負に発現」,「発現しない」の離散値に変換している．この離散化によって連続値の持っている情報が失われてしまう．決定木作成アルゴリズム C4.5 は連続値情報を扱うことが出来るので，この離散化によって失われた情報を補完するのに C4.5 を用いた．

決定木学習とは，属性によって特徴付けられた事例集合のどの属性で分類したら評価値が最適になるのかを計算によってノードが属性，葉ノードがクラスを割り当てるツリーを順次に生成し，その事例の属するクラスを判定する学習方法である．この学習によって得られた木構造によって表現された知識表現を決定木という．

たとえば，ある病院において患者の過去の症例がデータベース化されているとする．この過去のデータの蓄積から，どのような症状や健康状態の人に病気 A の病歴があるか否かを経験的に判定する知識を生成すれば，新たな患者が A を患っているか否かを判定するための助けになり便利である [22]．決定木とはこのような判定問題にしばしば利用される．図 4.6 の決定木では，条件部の属性として最低血圧 (数値)，血糖値 (+ か −)，コレステロール値 (+ か −) を考えて，最低血圧 (以下血圧) がある値以上か，もしくは他の値が + か − かでテストを行い，最終的に結論として生活習慣病である (◯) かあるいはそうではない (×) かを判定する．このとき，深さが浅く，頂点数も少ない決定木で良い判定ができれば理想的である [22]．この決定木を作成するツールはいくつかあるが，本研究では，

広く使われている C4.5 を用いた.

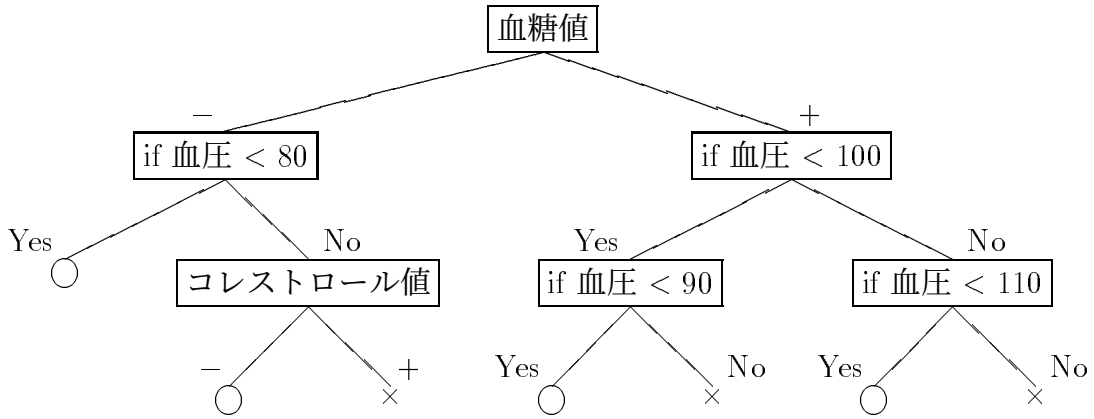

図 4.6: 決定木

4.6.1 決定木の生成

決定木学習は属性によって特徴付けられた事例集合から，ノードが属性，葉ノードがクラスに対応する木構造の知識表現を導くことである.

決定木の生成は次のように行われる．事例集合を D，事例を A_i，事例数を n，とすると，$D = \{A_1, A_2, \cdots, A_n\}$ と表現される．属性 A_i は属性値と呼ばれる要素 $a^i_1, a^i_2, \cdots, a^i{}_m$ を持つ集合として規定される．C4.5 の決定木生成は情報量を用いた $gain$ 法による．すなわち事例の部分集合を T，クラスを $C_1, C_2, \cdots, C_i$，その中でクラス C_j に分割される事例数を $freq(C_j, T)$ とすると，属性 A_i を分割属性に選んだ時の情報量の増加 $gain(A_i)$ は式 4.4 で計算される.

$$
\begin{aligned}
gain(A_i) &= info(T) - info_{A_i}(T) \qquad (4.4)\\
info(T) &= \sum_{j=1}^{i} \frac{freq(C_j, T)}{|T|} \log_2 \frac{freq(C_j, T)}{|T|}\\
info_{A_i}(T) &= \sum_{k=1}^{m_i} \frac{freq(a^i{}_k, T)}{|T|} info(a^i{}_k)
\end{aligned}
$$

C4.5 は，$gain(A_i)$ が最大となる属性を分割属性として選択することを部分木に対して繰り返し，トップダウン的に決定木を生成する [33].

4.6.2 決定木の問題

C4.5 は，各属性で分割したことによって得られる利得情報量を用いてルートノードからトップダウンで決定木を生成する手法であるが，利得情報量を用いた属性選択では，属性間に強い従属関係が存在する場合には，決定木の性能が著しく低下することが知られている [33].

第 5 章

プロトタイプシステムの構築

前章までの説明に基づき，本研究では，マイクロアレイデータから相関ルール発見等を行うプロトタイプシステムを構築した．本章では構築したシステムについて述べる．

5.1 システムの概要

本システムは Web 上で動作するシステムであり，ブラウザがあれば実行できる．ユーザからのデータ入力及び結果出力は全て Web 経由によって行われる．システムの実行はユーザが必要なパラメータを入力することにより行われる．入力されたパラメータは Web サーバ経由で CGI に受け渡される．CGI は，受け渡されたユーザのパラメータの指示に従って，データの加工，及びマイニングアルゴリズムの実行などを行い，その結果を Web サーバ経由でユーザ側のブラウザに表示する．また図 4.2 で示したようにデータマイニングの流れは，試行錯誤の繰り返しである．そのため，本システムでは必要な個所でプロセスの段階を戻ることが出来るシステム構成になっている．

本システムは大きく分けて相関ルール発見のデータを準備するデータ準備部分，相関ルール発見アルゴリズムのパラメータを設定し相関ルールを実際に実行する相関ルール発見部分，得られたルールを表示するルール表示部分，ルールの条件を満たす遺伝子を検索し表示する遺伝子検索部分および，得られたルールから決定木を作成する決定木部分の 5 つの部分から構成されている (図 5.1)．

本システムでは，出芽酵母細胞のマイクロアレイデータからマイニングを行うために，出芽酵母がもつ各遺伝子の特徴情報を YPD[1] データベースと ENZYME データベースから抽出し，マイニング用のデータに加え，データ補強をしている．YPD は出芽酵母の全

[1]Copyright (c) 2000 Proteome, Inc. All Rights Reserved. Not for distribution.

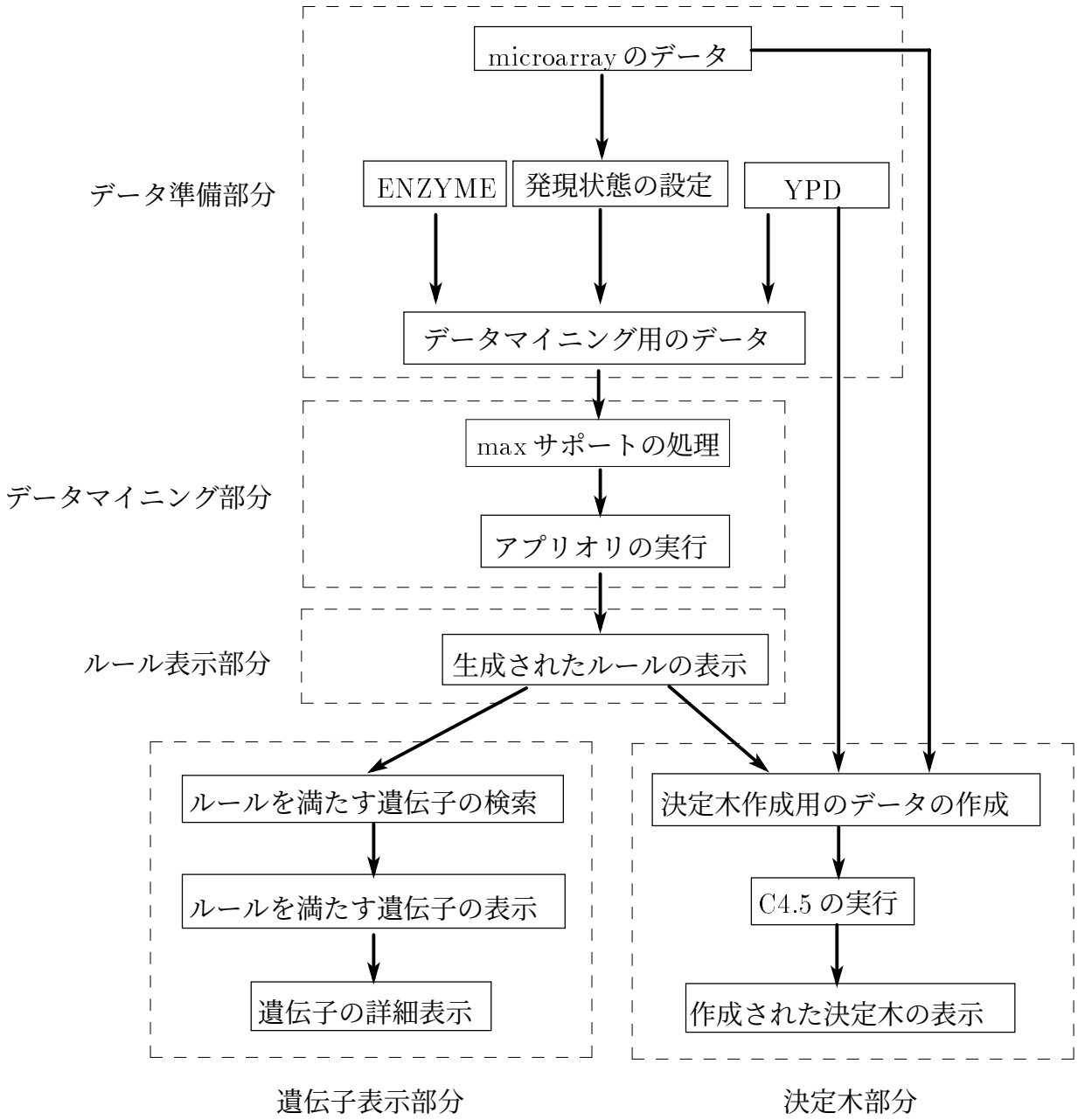

図 5.1: 本システムの概容

遺伝子の情報を網羅的に集めたデータベースであり，マイクロアレイデータと合わせるのに都合が良いからである．ENZYME はEC 番号による分類情報が使え，蛋白質の機能情報に関するデータソースとして有用なものであるからである．

相関ルール発見には，Magdeburg 大学の Christian Borgelt がフリーウェアとして公開しているプログラム apriori(Ver 2.5) を使用している [30]．決定木作成にはNew South Wales 大学の Ross Quinlan がフリーウェアで公開しているプログラム C4.5 を使用している [31]．本システムは，Perl，awk，C シェル，CGI プログラミングを用いて開発をおこなった．本システムの Web サーバはSUN Microsystems 社 Enterprise3000 で，OS はSunOS 5.6，メモリーは 1024Mbyte，CPU はUltraSPARC-II248MHz の 4CPU 構成である．

本システムは次のような手順で解析を行う．

1. ユーザが所定のパラメータをブラウザに入力すると，Web サーバ経由で CGI にパラメータが受け渡される．
2. CGI は，(1.) のパラメータに従ってマイクロアレイのデータを加工し，YPD と ENZYME データベースの情報を付加することで，マイニング用のデータを作成する．
3. CGI は，マイニング用データのトランザクション数をマイニングのパラメータとして提示する．
4. (3.) のマイニング用に準備されたデータのトランザクション数が不適切である，とユーザが考えれば，(1.) に戻りデータを作り直す．
5. ユーザは (4.) のトランザクション数を参考にマイニングを行うための適切なパラメータをブラウザに入力する．
6. CGI は，(5.) のパラメータに従ってマイニングを実行する．
7. CGI は，抽出したルールをユーザに示す．
8. ユーザは，抽出されたルールが不適切であると考えれば (6.) にもどり，マイニングのパラメータを設定し直す．
9. ユーザはさらにこのルールを次の 2 つの手法で分析することが出来る．
   - ルールの条件を満たす遺伝子集合を表示させる．
     さらに DBGET によって遺伝子の詳細情報を表示させる．
   - ルールに基づいて決定木作成用のデータを作成し，C4.5 で決定木を作成し，データの特徴付けを行う．

5.2 データ準備部分

この部分はマイニングを行うためのデータを生成する部分である．図5.3で示すブラウザの入力画面からパラメータを入力することで，ユーザは目的とするデータマイニングのデータを生成する．下にこの画面の設定項目の概要を説明する．

- マイクロアレイデータのうち測定エラーと思われるものを排除し，マイクロアレイデータの連続値を離散値に変換する．
- 機能既知遺伝子 (Well knwon gene)，部分的に機能が知られている遺伝子 (Partially knwon gene) 機能が全くわからない遺伝子 (Unknown gene) の内からどの組み合わせでデータマイニングするかを設定する．
- データ補強のためにENZYMEデータベースの情報を使う否か，EC番号をどのように使うかマイクロアレイのデータにどのように結合するのかなどを設定する．

この手順を行うことによって，図5.2のデータテーブルのデータが準備される．以下それぞれの設定項目について，詳細を説明する．

図 5.2: データマイニング用のデータテーブル

図 5.3: 設定画面

5.2.1 マイクロアレイデータの加工

第 3.3 章で述べたように，マイクロアレイデータは wild および mutant における遺伝子の発現量とその比から成る．この部分ではマイクロアレイのデータの遺伝子発現量を正に発現，負に発現，非発現の 3 つに分類することが出来る．まず，はじめに第 3.3 章の mutant と wild 及び mutant/wild の大きさを設定することでマイクロアレイの異常データの排除を行う．

第 3 章で示したように，マイクロアレイから得られるデータは連続量であるが，相関ルール発見手法は連続値を扱うことが出来ないので，マイクロアレイのデータは何らかの方法で離散化する必要がある．そこで，本システムではマイクロアレイの連続値を「正に発現」,「非発現」,「負に発現」の 3 値に変換して，データマイニング用のデータを準備することにした．本システムではデータ変換のためにデータの正規化を行う．図 3.1 で示したように，マイクロアレイのデータは対数正規分布確率密度関数に近似するので，マイクロアレイのデータの正規化は対数正規分布によって行う必要がある．これは，マイクロアレイ実験が発現量を蛍光物質の光として計測し，遺伝子の発現量としていることによる (溶液中の溶質の濃度の定量を行う分光分析において重要な役割を果たすランバート・ベールの法則が指数関数であることによる).

発現状態の設定はユーザが指定した $+\sigma$ によって行う．σ よりも正規化されたデータの標準偏差が大きければ「正に発現」，$-\sigma$ よりも小さければ「負に発現」，としてそれ以外は「非発現」としている．$\sigma = 1.3$ に設定したときのこの変換の概要を図 5.4 に示す．統計処理のルーチンは Perl の統計処理パッケージである Statistics::Descriptive(Ver2.4) を使用した．これは，基本的な記述統計関数を提供するモジュールであり，本システムでは標準偏差と平均値を求めるのに使用している．

5.2.2 遺伝子の分類と YPD の特徴情報の付加

ユーザは表 5.3 に示した遺伝子の分類を組み合わせた，組み合わせを示している表 5.4 の中から 1 つの組み合わせ選び，これをマイニング用のデータとする．ここでは，YPD と呼ばれるデータベースを使う．YPD とは Proteome 社が提供する出芽酵母の遺伝子・蛋白質データベースである[2] [54]．出芽酵母はモデル生物として古くから実験に使用されており，これまでに蓄積された分子遺伝学的な及び生物学的な情報は非常に広範囲にわたっている．YPD は，これらの情報を整理しまとめたものであり，遺伝子がコードする蛋白

[2]本研究ではアカデミックフリー版の YPD を使用している.

対数正規分布
による正規化
$\Longrightarrow$

図 5.4: マイクロアレイデータの加工

質の機能，細胞局在情報，相互作用情報[3] 文献情報などが網羅されている．アカデミックフリー版の YPD は 55 の項目があるが本研究ではその内の一部の項目のみを使った．

本システムでは，この YPD を用いて表 5.3 の遺伝子区分の判別を行った．これは YPD の表 5.2 の染色体番号以外の 4 つのフィールドを次のように用いて行った．Subcell localization，Molecular environment，Functional category，Cellular Role のうち全てのフィールドで unk(unknown を表す) 以外のアイテムがあれば，その遺伝子は Well known gene に分類する．4 つのフィールド内の 1 つでも unk 以外のアイテムがあれば Partially known gene に分類する．それ以外の遺伝子は Unknown gene に分類する．表 5.1 に YPD の一部と実際の分類を示す．第 1 フィールドが遺伝子分類であり，第 2 フィールド以降が YPD の情報である．そしてこの 4 つのフィールドのアイテムと染色体番号を YPD からの特徴情報として，マイクロアレイのデータに結合してマイニング用のデータとした．提供されたマイクロアレイのデータには遺伝子名がなく ORF 名のみであったので，ORF 名をキーとして結合した．

また YPD には，GenBank，SWISS-PROT，PIR というゲノムデータベースのが発行する登録番号 (アクセッションナンバー) も収録されている．表 5.1 の”GenBank Acc”，”SWISS-PROT Acc”，”PIR-INT Acc”のフィールドはそれぞれのデータベースのアクセッションナンバーを表している．

[3]商用版には付加されているが，アカデミックフリーでは削除されている．

表 5.1: YPD の一部と遺伝子分類

Type	ORF	Chr	GenBank Acc	SWISS-PROT Acc	PIR-INT Acc	Loc	Mol Env	Fun Grp	Role	···
unknown	YER106W	V	AAB64661.1	P40065	S50609			unk	unk	···
partially	YER113C	V	AAC03211.1	P40071	S50616	eds	int	unk	unk	···
known	YJR007W	X	CAA89529.1	P20459	A32108	cyt	psf	tlf	pro	···
partially	YHR139C	VIII	AAB68419.1	P13130	A28129	wal		str	unk	···
known	YLR170C	X		P35181		gol	per	str	sec	···
partially	YGL158W	VII	CAA96870.1	P38622	S47900			pki tra	unk	···
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

表 5.2: YPD のフィールド名

省略名	正式名	意味	アイテム数
Chr	Chromosome location of gene	染色体番号	16
Loc	Subcellular localization	細胞内局在	40
Mol	Molecular environment	環境種別	9
Fun	Functional category	機能分類	58
Role	Cellular Role	機能	48

表 5.3: 遺伝子の分類

遺伝子分類	意味	遺伝子数
Well known gene	機能既知遺伝子	1817
Partially known gene	部分的に機能が知られている遺伝子	1786
Unknown gene	機能未知遺伝子	2111

表 5.4: 遺伝子の分類パターン

パターン	Well known gene	Partially known gene	Unknown gene	gene set
1	◯	×	×	1817
2	×	◯	×	1786
3	◯	◯	×	3603
4	×	◯	◯	3897
5	×	×	◯	2111
6	◯	◯	◯	5714

5.2.3 ENZYME 情報の付加

ユーザは ENZYME データベースの情報を付加するか否かを選択する．付加する場合はその方法も選択する．ENZYME データベースは蛋白質の一種である酵素に関する情報を蓄積しているデータベースであり 2000 年 9 月 26 日現在，3829 エントリーであり，そのデータ量は 3.6MB である．ENZYME データベースの一部を付録 A に示す．

酵素は，触媒作用を持つ蛋白質であり，化学触媒反応と比べても数桁速い反応速度有し，かつ反応特異性がある物質である．酵素は古くから研究されており，多くの知見が蓄積されている．ENZYME データベースでは酵素の機能を EC 番号で系統的に分類をしている (図 5.7)．第 1 階層の分類を表 5.7 に示す．

YPD とマイクロアレイの情報を結合するには ORF 名をキーとして用いたが，ENZYME データベースには ORF 名の情報がないので，ORF 名をキーとして結合することが出来ない．そこで，次の 2 つの方法で結合した．

1 つめの方法は，ENZYME データベース内の GENES フィールドのサブフィールド SCE(出芽酵母の略称) の遺伝子名を用いた．先に述べたようにマイクロアレのデータ内には遺伝子名はない．そこで，遺伝子名と ORF が収録されている YPD を経由して結合することにした．すなわち，ENZYME 中の出芽酵母の遺伝子を YPD の遺伝子名と結合させることで，YPD の ORF に変換して，それをキーとしてマイクロアレイのデータと ENZYME の特徴情報を結合する (図 5.5)．これにより 625 個について ORF で結合を行うことが出来た．

2 つめの方法では，ゲノムネットの各種の分子生物学データベース間のリンク情報を蓄積している LinkDB を用いた．LinkDB は各種のデータベース間から引き出した直接リンク情報のほかに，直接リンクを逆向きにたどる逆引きリンクや，いくつかのリンクを経由

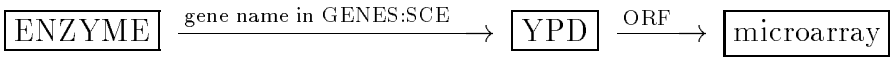

図 5.5: 遺伝子名による結合

してたどる間接的なリンクが入っている．表 5.5 に LinkDB の一部を示す．2000 年 9 月 26 日現在，ENZYME にリンクがある LinkDB のサブデータベースのエントリー数は 214546 エントリーである．データ量は 9.6MB である．YPD は商用のデータベースであり，ゲノムネットのデータベースではないので LinkDB 中には YPD へのリンク情報はない．そこで，YPD 中のアイテムの中で LinkDB 中にリンク情報が含まれている GenBank と SWISS-PROT，PIR のアクセッションナンバーに注目し，このアクセッションナンバーによって YPD と ENZYME を結合することを考えた．すなわち，ENZYME の情報を LinkDB のこれらのアクセッションナンバーによって YPD の ORF 名に変換してマイクロアレイに結合することが出来た (図 5.6)．この場合では，607 個の対応が取れた．表 5.6 に双方でどのくらい対応が取れたかを示す．GenBank のアクセッションナンバーでは YPD と LINKDB との一致が取れなかった．

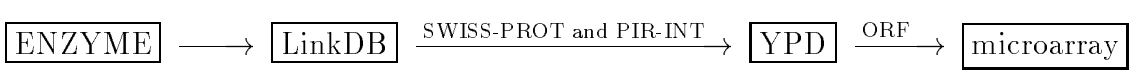

図 5.6: LinkDB による結合

ENZYME データベース中に含まれる情報のうち，遺伝子の特徴情報をよく表しており，かつ相関ルール発見のためにデータとして容易に使用できそうなものをデータマイニング用のデータに付加した．LinkDB と YPD とのアクセッションナンバーによる対応を表 5.8 に示す．

表 5.5: LinkDB のデータの一部

参照元	参照先	参照方式の種別	参照の経路 (間接参照の場合)
enzyme:1.1.1.1	genbank:M22342	indirect	enzyme→pir→genbank
enzyme:1.1.1.1	genbank:M24316	reverse	
enzyme:1.1.1.1	pdb:7ADH	original	
enzyme:1.1.1.1	pdb:8ADH	indirect	enzyme→swiss→pdb
enzyme:1.1.1.1	medline:93012919	indirect	enzyme→swiss→medline

表 5.6: LinkDB と YPD とのアクセッションナンバーによる対応

	YPD	対応が取れた数	LinkDB:ENZYME
GenBank ACC	6105	0	31190
SWISS-PROT ACC	4549	822(そのうち重複がないもの 771)	23926
PIR-INT ACC	5930	711	31192

表 5.7: EC 番号の分類

EC 番号 最上レベルの分類	種類
1	酸化還元酵素
2	転移酵素
3	加水分解酵素
4	除去付加酵素
5	異性化酵素
6	合成酵素

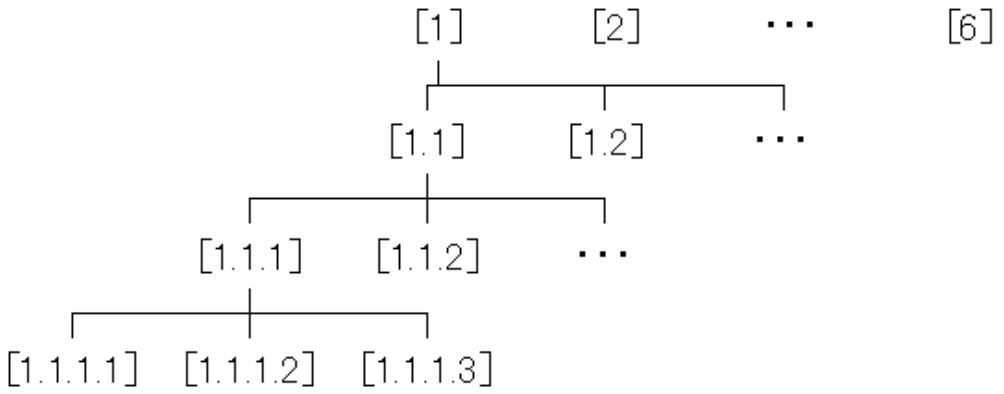

図 5.7: EC 番号の階層構造

表 5.8: ENZYME 内のエントリの種類と各アイテム数

エントリー	アイテムの種類
EC 番号 (第 1〜第 2 階層)	46
EC 番号 (第 1〜第 3 階層)	100
PRODUCT	484
COFACTOR	41
SUBSTRATE	561
PS(motif データベースへのクロスリファレンス)	415
EFFECTOR	10
PATH(この酵素が関係する反応経路)	84
MIM(OMIM データベースへのクロスリファレンス)	346
INHIBITOR	58

5.3 相関ルール発見部分

この部分では相関ルール発見のパラメータ設定を行い，実行させ，ルールを生成，表示を行う．図 5.8 で示すブラウザの入力画面からパラメータを入力して，実行することで図 5.9 で示すルールを得る事が出来る．この入力画面では，相関ルール発見のパラメータとして最小支持度，最大支持度，最小確信度の設定がある．ユーザはこれらのパラメータを試行錯誤して組み合わせながら有益なルール発見を行う．最小支持度と最大支持度のパラメータはトランザクション数で入力する．これは式 5.1 に従い，システムによって自動的に，パーセンテージ変換され，apriori プログラムに渡すパラメータとして使用される．このように，指定方法をパーセンテージではなくトランザクション数にしている理由は，本研究では小さな支持度を持つルールを生成することが多く，パーセンテージでは正確な指定がしにくいためである．

$$support = 100 \times \frac{(\text{トランザクションのサポート数} - 0.5)}{\text{データマイニング用に準備されたデータの全トランザクション数}} \tag{5.1}$$

図 5.8: 相関ルール設定画面


Loc:mem <- Mol:int Rol:tnc C 20 23 BG+(1.8%, 83.3%)
Loc:mem
Mol:int
Rol:tnc
Loc:mem <- Mol:int Rol:tnc Fun:tpt C 20 23 BG+(1.8%, 83.3%)
Loc:mem
Mol:int
Rol:tnc
Fun:tpt
Rol:tnc <- Loc:mem C 20 41 BG-(1.8%, 83.3%)
Rol:tnc
Loc:mem
Fun:set <- Mol:int Loc:mem C 20 41 BG-(1.8%, 83.3%)
Fun:set
Mol:int
Loc:mem
Loc:mem <- Rol:tnc Fun:tpt C 20 41 BG-(1.8%, 83.3%)
Loc:mem
Rol:tnc
Fun:tpt
Mol:dna <- Loc:nuc C 21 05 BG+ Rol:chr(1.8%, 83.3%)
Mol:dna
Loc:nuc
Rol:chr
Loc:mem <- Rol:tnc Fun:tpt C 20 23 BG+(1.8%, 83.3%)
Loc:mem
Rol:tnc
Fun:tpt
Loc:mit <- Mol:pol C 21 05 BG+(1.8%, 83.3%)
Loc:mit
Mol:pol
Loc:mem <- Mol:int Rol:tnc C 20 41 BG-(1.8%, 83.3%)
Loc:mem
Mol:int
Rol:tnc

図 5.9: ルール表示画面

5.3.1 ルール削除の工夫

相関ルール発見では一般的に，最小支持度を下げると生成されるルールが劇的に増加する．しかし本研究ではデータの性質と目的上，支持度がそれほど高くないルールに注目する必要があるので，どうしても生成されるルールが増えてしまう．そこで，本研究では現在 3 つの方法を用いて，あまり重要ではないと思われるルールを抑制している．

1 めの方法は，本研究で興味あるのがマイクロアレイに関連したルールのみにあるので，相関ルール発見で得られたルールのうちマイクロアレイに関係したルールだけを表示するようにした．これによって幾分か表示されるルールを削減することが出来た．

2 つめの方法はさらに，この最初の設定に加えて次のような工夫をした．相関ルール発見では、指定された最小支持度よりも高い支持度を持つラージアイテムセットを見つける処理を前半で行なう。仮に、支持度が 100 なアイテム（コンビニエンスストアの例で言えば、全ての客が購入する品物）があったとすると、原理的にこのアイテムは他の全てのアイテムと組み合わせが可能である。そこまで高い支持度でなくとも、指定された最小支持度に比べて非常に高い支持度を持つアイテムが多数ある場合、最小支持度を満足するラージアイテムセットの数が爆発的に増える。これを回避するため、本システムでは最大支持度という閾値をユーザが指定できるようにしている。この閾値は、もともと相関ルール発見の枠組には含まれていないものであるが、本研究のようにビジネスデータと異なる応用を行なう場合には、計算に支障を来すほど高い支持度を持つアイテムの存在が有り得る。これを回避するための簡便な方法として、最大支持度を用いてこのようなアイテムを最初から除外した後、アプリオリアルゴリズムを適用する。大雑把な方法ではあるが、これにより劇的にルール数を削減することができた。

しかし，まだ十分にルールを削除するには到らなかったので，3 つめ方法として apriori プログラムを第 4.5.3 章のルールの精錬で述べたモードで実行するオプションを使うことでルールを削除した．これによって，ようやくルールを大幅に削除することが出来た．

5.4 遺伝子検索部分

図 5.9 の相関ルール発見で得られたルールのアンカーをクリックすれば，データマイニング用に準備されたデータの中から，そのルールの条件を満たす遺伝子を検索し，図 5.10 のように遺伝子とそのアイテム群を表示する．このときルールの条件を満たすアイテムを赤字で強調表示する．遺伝子名をクリックすることによりゲノムネットの DBGET システムを通して，その遺伝子に関する詳細情報を表示することが出来る (図 5.11).

図 5.10: ルールを満足する遺伝子集合の表示画面

図 5.11: DBGET による遺伝子の詳細表示画面

5.5 決定木作成部分

5.5.1 概要

この部分では，相関ルールによって得られたルールのうち，ユーザが注目したルールに関連した決定木 C4.5 によって作成することで，得られたルールについてのさらに詳細な分析をおこなう．この利点として，次の 2 点を当初考えていた．

- C4.5 は連続値を扱うことができるので，離散値に変換したマイクロアレイの発現比を連続値に戻すことで，変換により失われた情報を補充されるので，より詳細な分析ができる．
- 相関ルール発見のデータ補強に使わなかった YPD の連続値情報を決定木作成用のデータに加えることで，データベースの情報を有効に活用することができる．これによって相関ルール発見の段階では活用できなかった情報を加えることで新たな知識が発見できる．

C4.5 は連続値を扱うことは出来るが，分類属性には連続値を設定することが出来ない．そのため，本システムではルール中の連続値以外のアイテムを目標属性に設定し，他のアイテムを説明属性に設定し，表 5.9 に示す YPD の連続値情報を加えることで，決定木作成用のデータを作成する．

決定木のデータに準備されるデータは相関ルール発見に使ったデータである．決定木作成用のデータは相関ルール発見用のデータから作成する．このときルールの条件を満たすトランザクションだけではなく，全トランザクションを作成の対象にする．これは，相関ルール発見によって得られたルールを分析するものであることによる．たとえば，パンを買う人はミルクを買う「ミルク $\Leftarrow$ パン」(=両方買う) というルールが，図 5.12 に示されるデータから得られたとする．このとき,「ミルク $\Leftarrow$ パン」を分析するには「ミルクだけ買った人」と「パンだけ買った人」も分析しなければ，ルールを分析したことにならない．言い換えれば，ラージアイテムセット {{ ミルク },{ パン },{ ミルク, パン }} の全てのアイテムに対して分析しなければ「ミルク $\Leftarrow$ パン」を分析したことにならないからである．

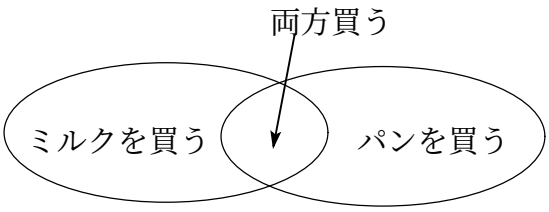

図 5.12: ベン図

図 5.13: 決定木表示画面

5.5.2 決定木の生成手順

この部分は次のような手順で解析を行う.

- **Step.1**
  ユーザは図 5.9 に表示されるルールのなかから決定木によって分析したいと考えるルールを選ぶ.

- **Step.2**
  選んだルールの下部にあるボタンを押すとそのボタンのラベルのアイテムが決定木のデータの分類属性に設定され，他のルール中のアイテムは説明属性に設定される. ただし，C4.5 は分類属性に連続値を扱うことは出来ないので，ルール中にマイクロアレイ発現状態のアイテムがあったとしても分類属性を設定するボタンは表示されない.

- **Step.3**
  分類属性として選択されたアイテムが相関ルール発見に用意されたトランザクションデータにおいてトランザクションにあれば True(1) を，なければ False(0) を決定木用のデータベースの分類属性フィールドに書き込む. マイクロアレイの発現状態を表すアイテム以外の説明属性に加えるアイテムについても同様な操作を行い，説明属性フィールドに書き込む. これを全てのトランザクションに対して行う.

- **Step.4**
  マイクロアレイの発現状態を表すアイテムは連続値に戻して C4.5 のデータに加える. このときルール中のマイクロアレイのアイテム「正に発現」および「負に発現」の状態は関係ない. 実験名のみをもとにして決定木作成用に準備するデータ全範囲に対して，連続値である発現比を説明属性として加える.

- **Step.5**
  YPD からは相関ルール発見で使えなかった数値情報であるデータ (表 5.9) を説明属性としてデータに加える.

- **Step.6**
  作成されたデータに対して C4.5 を使った決定木生成を行い，ユーザに表示する (図 5.13).

上で述べた Step.2 から Step.5 までの動作を図 5.14 に示す.

相関ルール発見に用意されたトランザクションデータ

	microarray						特徴情報		
	EXP1+	EXP1-	EXP2+	EXP2-	···	EXPn-	A	B	···
$gene_1$	1	0	0	1	···	0	1	0	···
$gene_2$	0	1	1	0	···	0	0	1	···
···	···	···	···	···	···	···	···	···	···
$gene_n$	1	0	0	1	···	0	1	1	⋱

⇓

「A ⇐ B, EXP1+」
A を分類属性として選択

⇓

「A⇐B,EXP1+」に対する決定木生成用のデータ

	相関ルールからの説明属性		分類属性	YPD から の説明属性	
	EXP1	B	A	a	···
$gene_1$	2.12	1	0	2.3	···
$gene_2$	1.11	0	1	1.76	···
···	···	···	···	···	···
$gene_n$	2.16	1	1	8.5	···

⇓

「EXP2- ⇐ EXP1+, A」
A を分類属性として選択

⇓

「EXP2-⇐EXP1+,A」に対する決定木生成用のデータ

	相関ルールからの説明属性		分類属性	YPD から の説明属性	
	EXP1	EXP2	A	a	···
$gene_1$	2.12	0.39	0	2.3	···
$gene_2$	1.11	3.23	1	1.76	···
···	···	···	···	···	···
$gene_n$	2.16	0.13	1	8.5	···

図 5.14: 決定木用のマイニングデータの作成

5.5.3 まとめ

相関ルール発見によって得られたルールを C4.5 に導入し結果を解析すると，ルール中のアイテム同士の属性間の影響が大きく，第 5.5.1 章で考えていたようには分析できなかった．しかしこの手法は十分に検討したわけではないので，有効ではないと結論付けることは出来ない．相関ルール発見によるルールを元に決定木導出を行う手法が提案されおり [34]，これらの関連研究を参考にして更なる検討が必要であると考える．

表 5.9: 追加した YPD のデータ

フィールド名	意味
Full_pi	Predicted isoelectric point of the full length protein
Mature_pi	Predicted isoelectric point of the mature form of the protein
pi_plus_one_plus	Isoelectric point calculated with one additional positive charge
pi_plus_one_minus	Isoelectric point calculated with one additional negative charge
MW_full	Molecular weight of the full length protein
MW_mature	Molecular weight of the mature form of the protein
Codon_Bias	Calculated by the method of Bennetzen and Hal
CAI	Codon Adaptation Index, calculated according to Sharp and Li
Intron	The number of spliceosomal introns within the gene
Full_len	Length (in amino acid residues) of the full length protein
Mat_len	Length (in amino acid residues) of the mature protein

第 6 章

遺伝子機能推定

6.1 概要

本研究ではマイクロアレイの遺伝子発現データから，相関ルール発見によって機能未知遺伝子の機能を推定するシステムの構築を行った．このシステムによって得られるルールから発現情報と遺伝子の機能を特徴付けるアイテム同士の関係を導き出すことは出来る．よって得られたルールを満たす遺伝子集合から遺伝子間の関係を得る事が出来る．だが，遺伝子同士の関係は直接は得る事が出来ない．すなわち，遺伝子機能推定を行うために遺伝子同士の直接の関係を求めることは出来ない．そこで，遺伝子集合同士から遺伝子間の関係を求めることで機能推定を行うことを試みた．本システムでは，出芽酵母の遺伝子を機能既知と機能未知を分けてマイニングすることが出来るので，遺伝子機能推定にはこの機能を用いて行った．

まず，機能既知および機能未知の 2 つの遺伝子群に対して網羅的にパラメータを変化させマイニングを行い，相関ルールを抽出した．その結果，双方から抽出されたルール集合の中には，少ないながらも共通するルールが含まれていることがわかった (図 6.1)．以下ではこのルールを共通ルールと呼ぶ．そしてルール集合の両方に存在する共通ルールの全体集合を共通ルール集合と呼ぶ．網羅的探索によってそれぞれの遺伝子群から得られたルール数および共通ルール数を付録 B に示す．この探索は，表 6.1 のパラメータを設定し，表 6.2 のようにパラメータを網羅的変化させて行った．なおこのとき，双方のマイニングで異なる影響が出ないように，最大支持度は使っていない．なお，この探索はブラウザ上からの操作ではなく，システムのエンジン部分だけを使用して，UNIX のシェルから直接行った．

表 6.1: パラメータ設定値 1

パラメータ項目	設定値
mutant	500
wild	500
mutant/wild	500
method	log
contain	all

表 6.2: パラメータ設定値 2

パラメータ項目	範囲	刻み幅
σ	1.3〜3.3	0.3
最小確信度	60〜100%	10%
最小支持度	5〜10 トランザクション	1 トランザクション

6.2 遺伝子集合同士の類似度計算

次に，先ほど網羅的に計算した中から共通ルールがある程度あり，かつ全ルールの組み合わせを計算するのに適当な計算時間で終了すると考えらるパラメーターセットを 1 つ (表 6.3) 選び，共通ルール集合を満足する機能既知/機能未知遺伝子集合が有意に似ているかを調べた．このとき得られた共通ルール集合を付録 C に示す．

図 6.2 に類似度計算の概要を示す．以下にその手順を説明する．共通ルールは機能未知遺伝子には発現状態を表すアイテムしかないので，共通ルールは発現状態を表すアイテムのみによって構成される．双方の遺伝子群から得られるルール集合の全ルールによって類似度を計算する．このとき，ルール中に出てくる特徴情報を表すアイテムを削除し，発現状態を表すアイテムのみによって類似度を計算する．これは，発現状態を表すアイテムが機能既知遺伝子群にしかなく機能未知遺伝子群には全くないので，類似度計算によって公平な比較ができないのを防ぐためである．具体的には次の様に行う．まず，各ルールについて，ルールの条件を満たす ORF を全て検索する．このとき ORF 名だけではなく，どのマイクロアレイ実験で ORF が発現したかを示すアイテムも ORF 名と同時に検索する．この，どのマイクロアレイ実験で ORF が発現したかを示すアイテムを発現状態アイテム

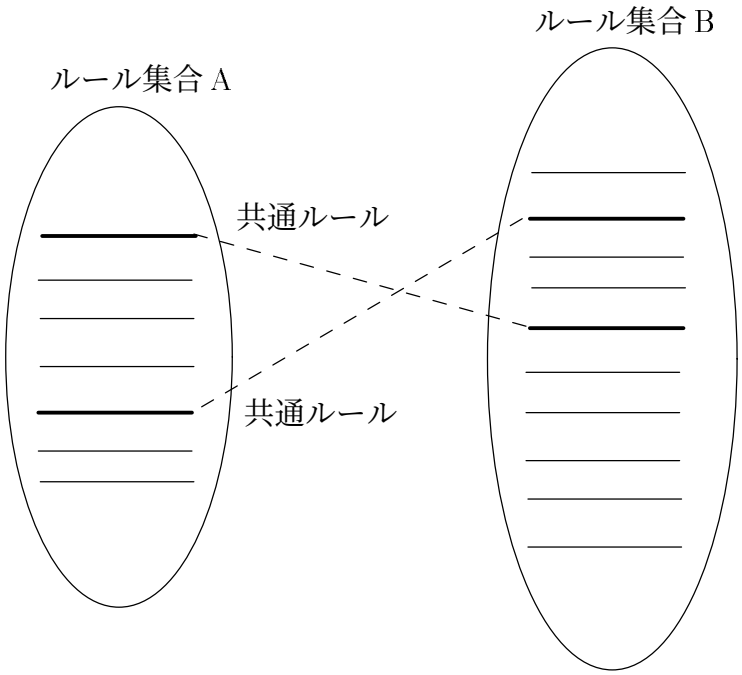

図 6.1: 共通ルール

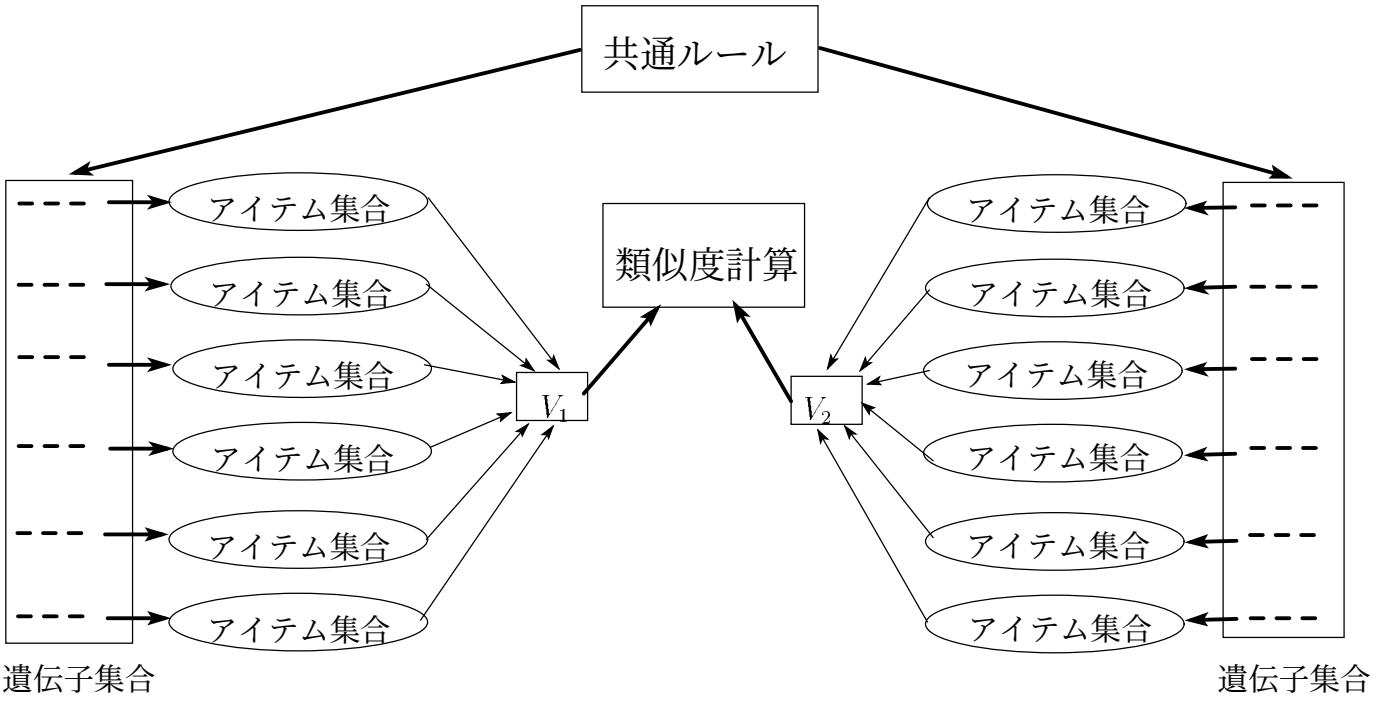

図 6.2: 類似度計算

と呼ぶことにする．1 つのルールから複数個の ORF と ORF ごとの発現状態アイテムを検索する．この結果を基にして，発現状態アイテムを要素にしたベクトルを 1 つ作る．このベクトルの要素の大きさは遺伝子群中に存在するそれぞれの発現状態アイテムの数である．すなわち，1 つのルールごとに 1 つのベクトルを作る．よってルールの個数だけベクトルが作られる．このルールの個数だけ作られたベクトルをベクトル集合と呼ぶ．

このようにして，機能既知遺伝子集合と機能未知遺伝子集合の双方のマイニングによって得られるそれぞれのルール集合に対してそれぞれのベクトル集合を作り出す．そして，双方のベクトル集合を全ての組み合わせによってベクトルの内積を式 6.1 で計算する．

$$\begin{aligned}
\overrightarrow{Known} &\equiv (Exp_1+, Exp_1-, Exp_2+, Exp_2-, \cdots, Exp_n+, Exp_n-) \\
\overrightarrow{Unknown} &\equiv (Exp_1+, Exp_1-, Exp_2+, Exp_2-, \cdots, Exp_n+, Exp_n-) \\
similarity(\overrightarrow{Known}, \overrightarrow{Unknown}) &\equiv \frac{\overrightarrow{Known} \cdot \overrightarrow{Unknown}}{|\overrightarrow{Known}||\overrightarrow{Unknown}|}
\end{aligned} \tag{6.1}$$

表 6.3: 類似度計算のパラメータ

パラメータ項目	設定値
σ	1.9
最小確信度	70%
最小支持度	5 トランザクション

全ルールを対象にして類似度計算したものを図 6.4 に，共通ルールのみを対象にして類似度を計算したものを図 6.3 に示す．なおこのとき類似度は%によって表示している．図 6.4 と図 6.3 より，共通ルールの類似性は全ルールの類似性よりも高くなっていることが示される．よって共通ルールと全ルールの間には明らかに有意な結果があるといえる．

6.3 共通ルールを満足する遺伝子集合

前節のパラメータでマイニングを行うと，29 個の共通ルールが得られる．このうちの 1 つの共通ルール (Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_45_ARM8.TXT+ Exp:C_43_21_ARM8.TXT+

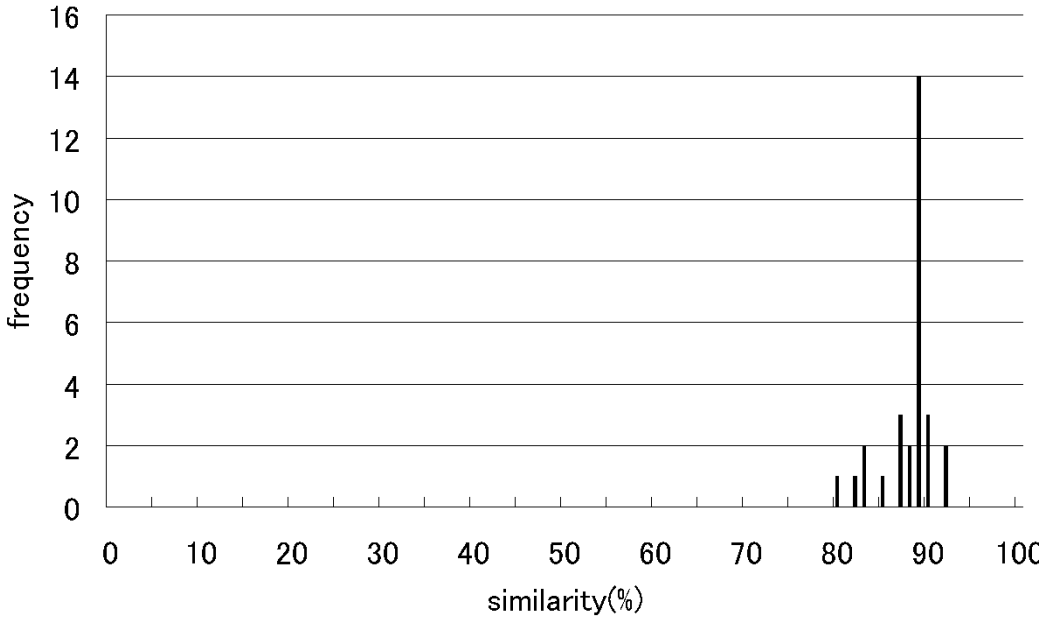

図 6.3: 共通ルールの類似度

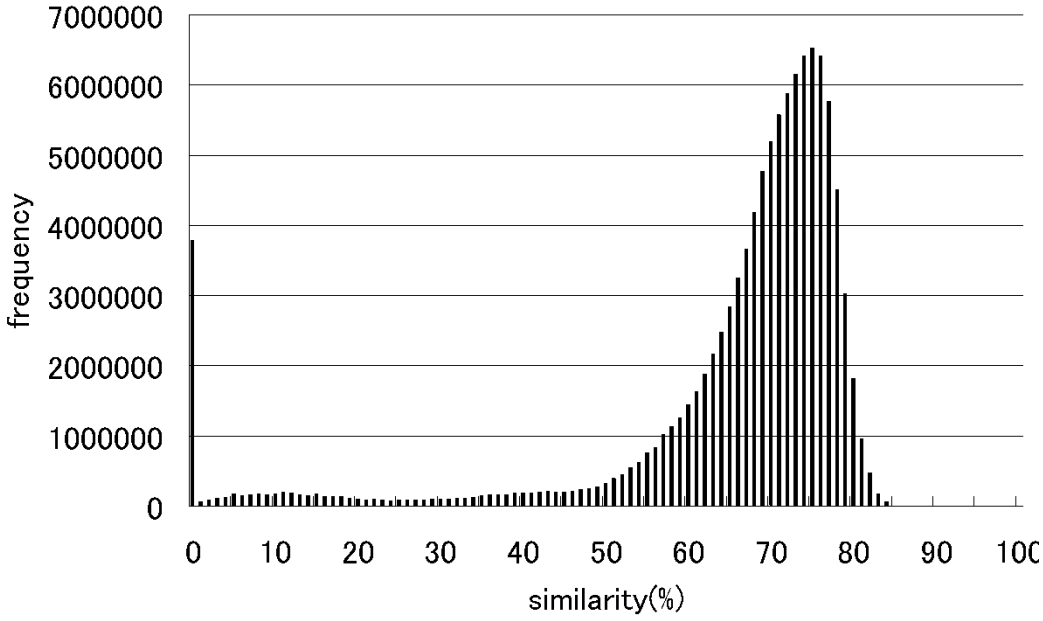

図 6.4: 全ルールの類似度

) を満足する機能既知および機能未知遺伝子集合の例を以下に示す．なお 29 個の共通ルールのそれぞれに対する遺伝子集後は付録 D に示す．一見して，rRNA(ribosomal RNA) 関連の遺伝子が多く含まれていることが分るが，生物学的にこれは普通の現象である．なぜならば rRNA は細胞内で最も豊富な RNA であり，さかんに増殖している細菌内において全体の 80%以上を占めるからである．これらの分子は，リボソームという蛋白質合成が起こる構造体の構成要素となっている [37].

```
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR181C RPS6B; 40S ribosomal protein S6e [SP:RS6_YEAST]
YDL191W RPL35A; 60S ribosomal protein L35e [SP:RL35_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL253W HXK2, HKB, HEX1; hexokinase II [EC:2.7.1.1] [SP:HXKB_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAS
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YJL138C TIF2, TIF41B; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAS
YKR094C RPL40B; 60S ribosomal protein L40e [SP:RL40_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR075W RPL10; 60S ribosomal protein L10e [SP:RL10_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YPL079W RPL21B; 60S ribosomal protein L21e [SP:R21B_YEAST]
unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOR129C unknown
```

第 7章

おわりに

7.1 まとめ

ゲノムサイズが短い微生物をはじめとして，各種モデル生物の完全長配列決定が多数報告される現在，ゲノム解析研究の焦点は配列決定から機能解析にシフトしつつある．マイクロアレイは，大量の遺伝子に関する発現情報を同時に測定できることから，ゲノムワイドな機能解析の道を開く新しい技術として，大いに注目されている．しかしながら，マイクロアレイ実験がもたらす大量の発現情報を如何に処理し，遺伝子機能の解析や遺伝子ネットワークの解明に結びつけるかについてはこれといって確立した手法がなく，各種の情報処理技術を試験的に投入することにより，有効な手法を模索しているのが現状である．

本研究では，ビジネスデータからの知識発見の分野で成功していると言われる相関ルール発見手法を用い，出芽酵母に関するマイクロアレイデータを解析することを試みた．相関ルール発見手法の枠組では，異種データを統合し，それらに跨る規則を発見することが容易に行なえる．この特徴を利用し，本研究ではマイクロアレイ実験の結果である遺伝子の発現量データに加えて，YPD および ENZYME という 2 種類のデータベースから得られる各種の機能情報や分類情報を導入し，これらのデータから相関ルールを発見するプロトタイプシステムを Web 上に構築した．ユーザはブラウザだけを用いてシステムを操作し，マイクロアレイデータの加工法や相関ルール発見の各種パラメータを指定し，試行錯誤しながら各種の解析を行なったり，ハイパーリンクを辿ってさらに詳細な検証を行なうことが可能である．データ加工に関しては，マイクロアレイデータの性質や相関ルール発見の枠組を考慮し，発現量比データの分布の正規化や,「正に発現する」「発現しない」「負に発現する」という 3 値への離散化を行なう際の閾値などについても指定を行なえる

ようにした．一方，離散化による情報の欠落を補うために，相関ルール発見の後処理として，特定のルールに関連して決定木生成を行なう機能も試験的に実装した．

本研究で構築したプロトタイプでは，出芽酵母の遺伝子全体を対象としたマイニングができることに加えて，ユーザの指定により，機能既知遺伝子や機能未知遺伝子に限定してマイニングを行なう機能も実装した．この機能を用いて，機能既知および未知の遺伝子に対し，同じパラメータを用いた相関ルール発見を行ない，全体の一部ではあるが共通ルールが得られる事，共通ルールを満足する各遺伝子集合の特徴情報に類似性があることなどを確認した．このことは，本システムにより遺伝子機能推定が行なえる事を直接的に意味するものではないが，今後さらなる検証を行なうことにより，遺伝子機能推定に有用な情報を得られる可能性があることを端的に示した．

7.2 今後の展望

プロトタイプシステムを構築する中で明らかになった問題と，現システムで改良及び追加を行う必要があると考えた点を，以下に述べる．

- 他のデータベースの追加
  現在のシステムでは，遺伝子機能情報が割合よく整理されている YPD データベースと蛋白質の一種である酵素の機能を系統的に分類している ENZYME データベースに蓄積されている情報をマイクロアレイのデータ補強に使ったが，これだけでは十分にデータ補強が出来ているとはいえない．出芽酵母はモデル生物として盛んに研究がなされており，他のデータベースにも多くの知見が蓄積されており，これらのデータベースからの情報を積極的に追加する必要がある．

- データ準備
  出芽酵母に存在する遺伝子の全てに遺伝子名が全てについているわけではない．本システムでは遺伝子名を LinkDB のリンク情報をもとに ORF 名に展開して遺伝子を一致させている．しかしながら現在は一部人手でこの処理を行っているため，手間がかかる．また，ゲノムネット以外のデータベースには LinkDB のリンク情報はないのでデータを一致させるのにはさらに手間がかかる．柔軟にこれらのデータベースを準備する手法の開発が今後の課題になると考える．また本研究では問題にならなかったが，出芽酵母には何種類かの株があり，同じ遺伝子でも呼び方が違うこともある．よって，整合性を取る必要が生じた場合には，対策が必要になるかもしれない．

- 離散値以外の情報の使用

  本研究では相関ルール発見が扱える情報として離散値だけを考えたため，マイクロアレイ実験の発現量比のデータを「正に発現している/いない」および「負に発現している/いない」という具合に 2 値データに変換して用いている．またデータ補強としてデータベースから加える情報は離散値情報に限定しており，データベースに蓄積されている自然言語情報，数値情報，塩基配列情報などの様々な情報を有効に活用しているとはいえない．たとえば，自然言語情報であれば文章の類似度を用いて分類し，この分類情報を使うとか，あるいは塩基配列情報の相同性などを用いるなどすればこれらの情報を離散値情報に変換してマイニングのデータにデータ加えることが出来る．他にも各種のデータ補強をすることにより有用な知識を発見することが可能と思われる．

- 最適なアプリオリのパラメーターの設定

  本システムはプロトタイプということもあり，アプリオリの最小支持度，最大支持度，最小確信度の組み合わせの検討はしていない．そのため適切なパラメータを設置しないと組み合わせの爆発が起こり，膨大なルールが生成されてしまったり，メモリの制約から最後まで計算を完了できないことがある．今後どのようなパラメーターをどのような範囲で設定すればよいかを検討する必要がある．

- ルール数の抑制

  本システムはパラメータによっては大量のルールが得られるが，冗長なルールも多く含まれている．そこで，本システムは冗長なルールの抑制の方法として次の 2 つの方法を行っている．1 つめは相関ルール発見手法におけるルール削減方法である．すなわち，ルールの頭部と本体の独立性に注目してルール数を抑制している．２つめは表示上の工夫で，マイクロアレイに関連するアイテムがあるルールのみを表示させている．どのような方法でルールを表示する．特に，前者のルールの抑制方法が生物学的な情報を損なっていないかどうか，また他のルール抑制手法の検討も必要である．

- 得られた相関ルールの分析手法

  本システムでは得られた相関ルールをさらに分析する 2 つの手法を提供している．1 つめの手法は，ルールを満たす遺伝子群を表示し，さらにゲノムネットのサービスである DBGET にリンクしている．2 つめは，得られた相関ルールについて決定木による分析を行うことができる．だが現行の方式では，相関ルール発見と決定木生成の相性が悪く，期待した分析は出来なかった．この方式によって相関ルールを

分析する必要があるかどうかは，データマイニング分野の関連研究をもとにして更なる検討が必要である．現行の方法では得られた相関ルールを十分に分析できるとは言いがたい．また相関ルール発見ではトランザクションデータベースにおいて出現頻度が少ないアイテムは適切に説明することが出来ない．そのため相関ルール発見の後処理は重要であると考えている．したがって，現行の分析方法検討と新しい分析手法を増やす必要があると考える．

- マイクロアレイデータの離散化方式の検討
  本研究では，連続値であるマイクロアレイから得られる発現量比を，ユーザーが設定したσによって「正に発現する/しない」,「負に発現する/しない」という2値に離散化している．この方法では連続値である発現量比の情報の損失が大きいので，より生物学的な知見に即した細かい離散化の方法を考えることが必要かもしれない．ただし，あまり細かく離散化を行うと相関ルール発見の性質を生かせなくなるので詳細に検討する必要がある．また，タイムコースのデータ及び生物学的な知見に基づいた構造をもつデータを加え，アプリオリアルゴリズム以外のアルゴリズムの検討と共に考える必要がある．

謝辞

本研究にあたり，終始御懇切なる御指導，御助言を賜りました，遺伝子知識システム論講座　佐藤賢二助教授に衷心から感謝の意を表します．佐藤賢二助教授の鋭い御指摘に，頭を抱えることも多々ありましたが，同時に身近な兄貴的な存在として，研究以外の事柄でも親身になって相談に乗って下さいました．

本研究に関して様々の御教授を頂いた遺伝子知識システム論講座　小長谷明彦教授に深く感謝致します．小長谷明彦教授は御多忙にも関わらず，基本的な内容の教授に時間を割いて下さいました．

サブテーマのデータマイニングに関して，熱心な御指導を賜った知識創造論講座 Tu Bao Ho 教授に感謝致します．サブテーマの研究の際に学んだ研究への姿勢は，本研究の遂行に不可欠なものでした．また，サブテーマにおいて得た知識が本研究のバックグラウンドになっています．．

遺伝子知識システム論講座　高橋勝利助手に深く感謝致します．高橋勝利助手には，数々の御助言と適切な御指導を頂きました．

マイクロアレイのデータを御提供下さった九州大学大学院　久原哲教授に深く感謝致します．久原哲教授にはデータ解析の際に貴重なご助言も賜りました．久原哲教授の御協力なくしてこの研究は遂行不可能でした．

本システム設計にあたり，貴重な御助言を頂いた福岡国際大学の古市恵美子助教授に感謝致します．

本研究遂行の過程で様々のご協力を頂いた遺伝システム論講座の諸氏および九州大学大学院の久原研究室の皆様に感謝致します．

直接的，間接的に本研究に助言，示唆を与えて下さった知識科学研究科の教官および学生に感謝致します．

参考文献

[1] KEGG Web Page, http://www.genome.ad.jp/kegg/

[2] Genome Net Web Page, http://www.genome.ad.jp/

[3] 高木利久, 金久實 編：ゲノムネットのデータベース利用法 [第 2 版], 共立出版出版株式会社,1998.

[4] 柳田充弘：酵母, 共立出版出版株式会社,1996.

[5] 金久實 担当編集委員：ヒューマンゲノム計画, 共立出版出版株式会社,1997.

[6] Oliver, S.G.: From DNA sequence to biological function. Nature 379, 597-600 (1996).

[7] DeRisi JL, lyer VR, Brown PO : Exploring the metabolic and genetic contorl of gene expression on genomic scale, Science, Vol.278, pp.680-686, 1997.

[8] Chu S, DeRisi JL, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I : The transcriptional program of sporulation in budding yeast, Science, Vol.282, pp.699-705, 1998.

[9] The Brown Lab Web Page, http://cmgm.stanford.edu/pbrown/

[10] Schena M, Shalon D, Davis RW, Brown PO : Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Scienece, Vol.270, pp.467-470, 1995.

[11] 村松正明, 那波宏之 監修：DNA マイクロアレイと最新 PCR 法, 秀順社, 2000.

[12] Help! The data are coming, (opinion), Nature, Vol.399, pp.505, 1999.

[13] S.Schulze-Kremer : Discovery in the human genome project, Comm. ACM, Vol.42, pp.62-64, 1999.

[14] 金久實：ゲノムネット, bit,Vol.31, No.8, 共立出版株式会社,1999.

[15] 西尾章次郎： 大規模データベースにおける知識獲得, 情報処理学会誌, Vol.34,No.3, pp343-350,1993.

[16] 河野浩之, 西尾章次郎, Jiawei Han :データベースからの知識獲得技術, 人工知能学会誌, Vol.10-No.1, pp38-44,1994.

[17] 大久保公作, 森下真一： 発現情報データベースとクラスタリング, bit,Vol.31-No.12, 共立出版株式会社,1999.

[18] Agrawal, R. and Srikant, R. : Fast Alogrithms for Mining Association Rules, Proc. VLDB, pp.487-499, 1994.

[19] 特集 大規模データベースからの知識獲得，人工知能学会誌, Vol.12-No.4,1997.

[20] ゲノムサイエンス, 共立出版株式会社, 1997.

[21] J.R. キンラン, 古川康一 監訳：AI によるデータ解析, 株式会社トッパン,1995.

[22] 福田剛志，森本康彦，森下真一，徳山豪 :データマイニングの最新動向，情報処理論文誌，Vol.37，No.7, 1996.

[23] Pieter,Adriaans.Dolf,Zantinge., 山本英子, 梅村恭司訳： データマイニング, 共立出版株式会社,1998.

[24] マイケル J.A. ベリー，ゴード・リノフ：SAS インスティチュート ジャパン/江原淳, 佐藤栄作 共訳:データマイニング手法, 海文堂出版株式会社,1999.

[25] P. キャベナ, P. ハジニアン, R. スタッドラー, J. ベルフィース, A. ザナシー著日本アイ・ビー・エム株式会社/河村住洋, 福田剛志監訳, 日本アイ・ビー・エム株式会社/ナショナル・ランゲージ・サポート訳： データマイニング活用ガイド, エスアイビー・アクセス, 株式会社星雲社, 1999.

[26] 福田剛志，森下真一：相関ルールの可視化と重要ルールのふるい分け「相関ルールの可視化について」電子情報通信学会 技術研究報告，Vol.95, No.81, pp.41-48, May, 1995

[27] 布施田敏樹：ゲノムデータベースにおける柔軟なデータ加工およびマイニングシステムの構築に関する研究, 北陸先端科学技術大学院大学 知識科学研究科 修士論文, 2000.

[28] 森下真一：情報検索からデータマイニングへ，電子情報通信学会，Vol.97,No.10, pp.1030-1032.

[29] データウェアハウス・レポート'98, 株式会社コンピュータ・エイジ社，1998.

[30] Web Page, http://fuzzy.cs.uni-magdeburg.de/ borgelt/

[31] Web Page, http://www.cse.unsw.edu.au/ quinlan/

[32] 櫛雄介, 稲積宏誠：複数属性に注目した決定木生成に関する検討, 人工知能学会全国大会論文集,1999.

[33] 吉澤有美, 稲積宏誠：論理最小化による決定木生成, 計測自動制御学会知能システムシンポジウム資料, pp.51-56,1999.

[34] 寺邊正大，片井修，椹木哲夫，鷲尾隆，元田浩：相関ルールにもとづく属性生成手法，人工知能学会誌，Vol.19, No.1, 2000.

[35] 星田昌記：遺伝子情報処理への挑戦, 出版株式会社,1994.

[36] 小長谷明彦：遺伝子とコンピュータ, 共立出版出版株式会社,2000.

[37] T.A.Brown 著, 村松正實 監訳：ゲノム, メディカル・サイエンス・インターナショナル,2000.

[38] 小関治男, 永田俊夫, 松代愛三, 由良隆 著：分子生物学, 化学同人, 1996.

[39] 九州大学大学院生物資源環境科学研究科 遺伝子資源工学専攻遺伝子制御学講座ホームページ http://www.grt.kyushu-u.ac.jp/grt-docs/mogt/

[40] 白髭克彦, マイクロアレイが開く酵母研究の新時代, 実験医学,Vol.17, No.19, pp.2544-2549, 1999.

[41] 大久保公策, ゲノムワイドな遺伝子発現情報の収集と解析, 実験医学,Vol.17, No.19, pp.2537-2543, 羊土社, 1999.

[42] 油谷浩幸，ヒト遺伝子解析と DNA チップ，遺伝，Vol.54，No.4，裳華房，2000.

[43] 久原哲，田代康介, 牟田滋：DNA チップの情報科学的取り扱い, No.432, 1999.

[44] 村上康文：モデル生物ゲノム解析研究の展開，GDB ニュースレター，第２号，1997.

[45] NCBI WEB Page, http://www.ncbi.nlm.nih.gov/

[46] 阿久津達也：遺伝子ネットワークの推定アルゴリズム, 数理科学, No.432, 1999.

[47] 中谷明弘, 森下真一：解説ナノレベルのゲノム実験室, 情報処理,Vol.40-No.3,pp.320-325, 1999.

[48] 中村祐輔研究室：DNA マイクロアレイー・DNA チップ技術, 2000.

[49] 中村祐輔研究室：ラボマニュアル～マイクロアレイ編, 2000.

[50] DNA マイクロアレイ実戦マニュアル，羊土社，2000.

[51] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. : Molecular portraits of human breast tumours, Nature, Vol.406, pp747-752.

[52] 辻本豪三, 田中利男 (編)，ゲノム機能研究プロトコール, 羊土社, 2000.

[53] Eisen, M.B. et al. : Proc. Natl Acad. Sci. USA, 95 : pp14863-14969, 1998.

[54] YPD Web Page, http://www.proteome.com/databases/index.html

[55] Goto, S., Nishioka, T., and Kanehisa, M. : LIGAND database for enzymes, compounds, and reactions, Nucleic Acids Res., 27, pp.377-379, 1999.

[56] Opinion Web Page, http://infofarm.cc.affrc.go.jp/ domon/opinion/opinion.htm

研究業績

Takahiro Naitou, Kenji Satou, Emiko Furuichi, Satoru Kuhara and Toshihisa Takagi: A System for Finding Association Rules from Microarray Data and Public Database, Genome Informatics 2000, UNIVERSAL ACADEMY PRESS, INC. TOKYO, JAPAN, pp.356-357.

第 A 章

付録

ENZYMEのエントリ

```
ENTRY       EC 1.1.1.1
NAME        Alcohol dehydrogenase
            Aldehyde reductase
CLASS       Oxidoreductases
            Acting on the CH-OH group of donors
            With NAD+ or NADP+ as acceptor
SYSNAME     Alcohol:NAD+ oxidoreductase
REACTION    Alcohol + NAD+ = Aldehyde or Ketone + NADH
SUBSTRATE   NAD+
            Primary alcohol
            Secondary alcohol
            Cyclic secondary alcohol
            Hemiacetal
PRODUCT     Aldehyde
            Ketone
            NADH
COFACTOR    Zinc
COMMENT     A zinc protein. Acts on primary or secondary alcohols or
            hemiacetals; the animal, but not the yeast, enzyme acts also on
            cyclic secondary alcohols
            The insect enzyme is a member of the nonmetallo-short-chain
            alcohol dehydrogenase (ADH) family (Proc.Natl.Acad.Sci.USA(1991)
            88, 10064-10068).
PATHWAY     PATH: MAP00010  Glycolysis / Gluconeogenesis
            PATH: MAP00071  Fatty acid metabolism
            PATH: MAP00120  Bile acid biosynthesis
            PATH: MAP00350  Tyrosine metabolism
```

```
            PATH: MAP00561  Glycerolipid metabolism
GENES       ECO: b0356(adhC) b1241(adhE) b1478(adhP) b3589(yiaY)
            HIN: HI0185(adhC)
            VCH: VC2033
            NME: NMB0546 NMB1304
            NMA: NMA0725(adhA) NMA1518(adhC)
            BSU: adhB gbsB
            MTU: Rv0761c(adhB) Rv1862(adhA) Rv2259(adhE2)
            AAE: aq_1240(adh2) aq_1362(adh1)
            TMA: TM0111 TM0920
            AFU: AF0024 AF0339 AF2019 AF2101
            PHO: PH0743
            PAB: PAB1511
            APE: APE1245 APE1557 APE1963 APE2239
            SCE: YBR145W(ADH5) YDL168W(SFA1) YGL256W(ADH4) YMR083W(ADH3)
                 YMR303C(ADH2) YOL086C(ADH1)
            CEL: K12G11.3 K12G11.4
            DME: CG3425(T3dh) CG3481(Adh) CG6598(Fdh)
            MMU: 1098256(Daq1) 87921(Adh1) 87926(Adh3) 87929(Adh5)
            HSA: 124(ADH1) 125(ADH2) 126(ADH3) 127(ADH4) 130(ADH6) 131(ADH7)
DISEASE     MIM: 103700  Alcohol dehydrogenase (class I), alpha polypeptide
            MIM: 103720  Alcohol dehydrogenase (class I), beta polypeptide
            MIM: 103730  Alcohol dehydrogenase (class I), gamma polypeptide
            MIM: 103740  Alcohol dehydrogenase (class II), pi polypeptide
            MIM: 600086  Alcohol dehydrogenase-7
MOTIF       PS: PS00059  G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVSAC]
            PS: PS00060  [GSW]-x-[LIVTSACD]-[GH]-x(2)-[GSAE]-[GSHYQ]-x-[LIVTP]-
                         [GAST]-[GAS]-x(3)-[LIVMT]-x-[HNS]-[GA]-x-[GTAC]
            PS: PS00061  [LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-
                         {PC}-[SAGFYR]-[LIVMSTAGD]-x(2)-[LIVMFYW]-x(3)-
                         [LIVMFYWGAPTHQ]-[GSACQRHM]
            PS: PS00913  [STALIV]-[LIVF]-x-[DE]-x(6,7)-P-x(4)-[ALIV]-x-[GST]-
                         x(2)-D-[TAIVM]-[LIVMF]-x(4)-E
STRUCTURES  PDB: 1A4U  1A71  1A72  1ADB  1ADC  1ADF  1ADG  1AGN  1AXE  1AXG
                 1BTO  1CDO  1D1S  1D1T  1DEH  1HDX  1HDY  1HDZ  1HLD  1HTB
                 1LDE  1LDY  1TEH  2OHX  2OXI  3BTO  3HUD  5ADH  6ADH  7ADH
DBLINKS     ExPASy - ENZYME nomenclature database: 1.1.1.1
            WIT (What Is There) Metabolic Reconstruction: 1.1.1.1
            BRENDA, the Enzyme Database: 1.1.1.1
            SCOP (Structural Classification of Proteins): 1.1.1.1
///
```

第 B 章

付録

パラメータとルール数の関係

表 B.1: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
1.3	60	5	10661	31443	87
1.3	60	6	9205	21115	54
1.3	60	7	8434	21115	53
1.3	60	8	8104	12102	34
1.3	60	9	7862	6024	18
1.3	60	10	7608	3126	8
1.3	70	5	5208	20552	6
1.3	70	6	4173	12981	1
1.3	70	7	3666	12981	1
1.3	70	8	3424	6089	1
1.3	70	9	3277	3140	1
1.3	70	10	3182	1243	1
1.3	80	5	3614	13671	2
1.3	80	6	3040	6100	0
1.3	80	7	2714	6100	0
1.3	80	8	2592	2279	0
1.3	80	9	2481	901	0
1.3	80	10	2400	525	0
1.3	90	5	2302	5003	0
1.3	90	6	2081	1734	0
1.3	90	7	1911	1734	0
1.3	90	8	1854	509	0
1.3	90	9	1803	183	0
1.3	90	10	1752	95	0
1.3	100	5	1292	4945	0
1.3	100	6	1071	1676	0
1.3	100	7	901	1676	0
1.3	100	8	844	451	0
1.3	100	9	793	125	0
1.3	100	10	762	37	0

表 B.2: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
1.6	60	5	14035	45442	254
1.6	60	6	13338	15991	113
1.6	60	7	12738	15991	112
1.6	60	8	12336	5504	43
1.6	60	9	11911	5504	43
1.6	60	10	11561	1662	17
1.6	70	5	5268	33186	26
1.6	70	6	4804	10532	8
1.6	70	7	4532	10532	8
1.6	70	8	4260	3731	2
1.6	70	9	4040	3731	2
1.6	70	10	3901	951	1
1.6	80	5	3121	18847	1
1.6	80	6	2825	8756	0
1.6	80	7	2667	8756	0
1.6	80	8	2558	1955	0
1.6	80	9	2431	1955	0
1.6	80	10	2340	432	0
1.6	90	5	2070	13232	0
1.6	90	6	1914	3141	0
1.6	90	7	1811	3141	0
1.6	90	8	1756	596	0
1.6	90	9	1719	596	0
1.6	90	10	1672	113	0
1.6	100	5	1119	13228	0
1.6	100	6	963	3137	0
1.6	100	7	860	3137	0
1.6	100	8	805	592	0
1.6	100	9	768	592	0
1.6	100	10	749	109	0

表 B.3: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
1.9	60	5	18236	17360	218
1.9	60	6	17137	17360	218
1.9	60	7	16036	2694	48
1.9	60	8	15089	2694	48
1.9	60	9	14013	2694	48
1.9	60	10	12944	457	5
1.9	70	5	7813	13535	29
1.9	70	6	6959	13535	29
1.9	70	7	6407	1962	6
1.9	70	8	5755	1962	6
1.9	70	9	5227	1962	6
1.9	70	10	4790	377	0
1.9	80	5	3868	7866	4
1.9	80	6	3283	7866	4
1.9	80	7	2965	1843	1
1.9	80	8	2771	1843	1
1.9	80	9	2472	1843	1
1.9	80	10	2222	258	0
1.9	90	5	2117	6804	2
1.9	90	6	1809	6804	2
1.9	90	7	1663	781	0
1.9	90	8	1585	781	0
1.9	90	9	1542	781	0
1.9	90	10	1431	90	0
1.9	100	5	1233	6804	0
1.9	100	6	925	6804	0
1.9	100	7	779	781	0
1.9	100	8	701	781	0
1.9	100	9	658	781	0
1.9	100	10	627	90	0

表 B.4: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
2.2	60	5	25781	49852	689
2.2	60	6	21545	3748	71
2.2	60	7	16929	3748	65
2.2	60	8	13593	3748	61
2.2	60	9	10272	157	5
2.2	60	10	7390	157	4
2.2	70	5	13185	34835	227
2.2	70	6	10152	3435	25
2.2	70	7	8129	3435	21
2.2	70	8	5842	3435	19
2.2	70	9	4226	138	2
2.2	70	10	3303	138	2
2.2	80	5	6115	33788	73
2.2	80	6	4393	2388	6
2.2	80	7	3324	2388	5
2.2	80	8	2762	2388	4
2.2	80	9	1929	133	1
2.2	80	10	1561	133	1
2.2	90	5	2194	33732	17
2.2	90	6	1652	2332	2
2.2	90	7	1332	2332	2
2.2	90	8	1195	2332	2
2.2	90	9	1114	77	0
2.2	90	10	960	77	0
2.2	100	6	998	2332	0
2.2	100	7	678	2332	0
2.2	100	8	541	2332	0
2.2	100	9	460	77	0
2.2	100	10	422	77	0

表 B.5: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
2.5	60	5	19467	13683	263
2.5	60	6	11679	13683	181
2.5	60	7	6729	428	7
2.5	60	8	4273	428	5
2.5	60	9	2798	428	2
2.5	60	10	1938	10	0
2.5	70	5	12195	11048	132
2.5	70	6	6506	11048	71
2.5	70	7	3805	377	3
2.5	70	8	2087	377	2
2.5	70	9	1412	377	1
2.5	70	10	1160	7	0
2.5	80	5	6082	10981	60
2.5	80	6	2954	10981	27
2.5	80	7	1679	310	0
2.5	80	8	1243	310	0
2.5	80	9	914	310	0
2.5	80	10	771	7	0
2.5	90	5	2084	10978	19
2.5	90	6	1097	10978	10
2.5	90	7	738	307	0
2.5	90	8	635	307	0
2.5	90	9	596	307	0
2.5	90	10	552	4	0
2.5	100	5	1752	10978	19
2.5	100	6	765	10978	10
2.5	100	7	406	307	0
2.5	100	8	303	307	0
2.5	100	9	264	307	0
2.5	100	10	254	4	0

表 B.6: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
2.8	60	5	4551	2026	30
2.8	60	6	2473	2026	8
2.8	60	7	1445	74	0
2.8	60	8	1071	74	0
2.8	60	9	884	74	0
2.8	60	10	790	1	0
2.8	70	5	2902	1613	12
2.8	70	6	1501	1613	2
2.8	70	7	1069	63	0
2.8	70	8	806	63	0
2.8	70	9	685	63	0
2.8	70	10	634	1	0
2.8	80	5	1640	1601	5
2.8	80	6	906	1601	0
2.8	80	7	692	51	0
2.8	80	8	635	51	0
2.8	80	9	573	51	0
2.8	80	10	538	1	0
2.8	90	5	765	1600	2
2.8	90	6	511	1600	0
2.8	90	7	448	50	0
2.8	90	8	433	50	0
2.8	90	9	418	50	0
2.8	90	10	405	0	0
2.8	100	5	548	1600	2
2.8	100	6	294	1600	0
2.8	100	7	231	50	0
2.8	100	8	216	50	0
2.8	100	9	201	50	0
2.8	100	10	193	0	0

表 B.7: パラメータとルール数の関係

σ	最小確信度	最小支持度	既知遺伝子のルール数	未知遺伝子のルール数	共通ルール数
3.1	60	5	1067	408	0
3.1	60	6	876	408	0
3.1	60	7	776	5	0
3.1	60	8	744	5	0
3.1	60	9	712	5	0
3.1	60	10	685	5	0
3.1	70	5	860	91231	10
3.1	70	6	717	349	0
3.1	70	7	670	4	0
3.1	70	8	641	4	0
3.1	70	9	618	4	0
3.1	70	10	600	4	0
3.1	80	5	695	10981	1
3.1	80	6	594	347	0
3.1	80	7	561	2	0
3.1	80	8	549	2	0
3.1	80	9	532	2	0
3.1	80	10	517	2	0
3.1	90	5	477	347	0
3.1	90	6	434	347	0
3.1	90	7	413	2	0
3.1	90	8	407	2	0
3.1	90	9	397	2	0
3.1	90	10	387	2	0
3.1	100	5	272	347	0
3.1	100	6	229	347	0
3.1	100	7	208	2	0
3.1	100	8	202	2	0
3.1	100	9	192	2	0
3.1	100	10	185	2	0

第 C 章

付録

共通ルール

表 C.1: 共通ルール

ルール番号	共通ルール
1	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_45_ARM8.TXT+ Exp:C_43_21_ARM8.TXT+
2	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_24_ARM8.TXT+ Exp:C_38_18_BG.TXT+
3	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_19_ARM8.TXT+ Exp:C_42_44_ARM8.TXT+
4	Exp:C_41_33_ARM8.TXT+ ← Exp:C_43_21_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
5	Exp:C_38_19_BG.TXT+ ← Exp:C_41_45_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
6	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_44_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
7	Exp:C_41_22_ARM8.TXT+ ← Exp:C_41_34_ARM8.TXT+ Exp:C_41_40_ARM8.TXT+
8	Exp:C_41_33_ARM8.TXT+ ← Exp:C_42_44_ARM8.TXT+ Exp:C_13_11_BG.TXT+
9	Exp:C_41_33_ARM8.TXT+ ← Exp:C_43_34_ARM8.TXT+ Exp:C_38_18_BG.TXT+
10	Exp:C_41_44_ARM8.TXT+ ← Exp:C_43_35_ARM8.TXT+ Exp:C_38_19_BG.TXT+
11	Exp:C_41_45_ARM8.TXT+ ← Exp:C_41_38_ARM8.TXT+ Exp:C_43_19_ARM8.TXT+
12	Exp:C_44_22_ARM8.TXT+ ← Exp:C_44_20_ARM8.TXT+ Exp:C_43_17_ARM8.TXT+
13	Exp:C_38_18_BG.TXT+ ← Exp:C_43_40_ARM8.TXT+ Exp:C_38_43_ARM8.TXT+
14	Exp:C_41_24_ARM8.TXT+ ← Exp:C_43_45_ARM8.TXT+ Exp:C_41_18_ARM8.TXT+
15	Exp:C_41_44_ARM8.TXT+ ← Exp:C_44_17_ARM8.TXT+ Exp:C_20_21_BG.TXT+
16	Exp:C_44_17_ARM8.TXT+ ← Exp:C_43_41_ARM8.TXT+ Exp:C_43_23_ARM8.TXT+
17	Exp:C_44_18_ARM8.TXT+ ← Exp:C_41_24_ARM8.TXT+ Exp:C_43_35_ARM8.TXT+
18	Exp:C_41_24_ARM8.TXT+ ← Exp:C_42_12_ARM.TXT+ Exp:C_43_34_ARM8.TXT+
19	Exp:C_41_47_ARM.TXT+ ← Exp:C_42_12_ARM.TXT+ Exp:C_38_44_ARM8.TXT+
20	Exp:C_41_19_ARM8.TXT+ ← Exp:C_41_44_ARM8.TXT+ Exp:C_43_42_ARM8.TXT+
21	Exp:C_41_26_ARM8.TXT+ ← Exp:C_44_01_ARM8.TXT+ Exp:C_38_18_BG.TXT+
22	Exp:C_41_33_ARM8.TXT+ ← Exp:C_43_34_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
23	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_45_ARM8.TXT+ Exp:C_38_18_BG.TXT+
24	Exp:C_41_33_ARM8.TXT+ ← Exp:C_41_23_ARM8.TXT+ Exp:C_38_18_BG.TXT+
25	Exp:C_44_09_ARM8.TXT+ ← Exp:C_44_15_ARM8.TXT+ Exp:C_44_03_ARM8.TXT+
26	Exp:C_44_07_ARM8.TXT+ ← Exp:C_44_21_ARM8.TXT+ Exp:C_43_19_ARM8.TXT+
27	Exp:C_41_33_ARM8.TXT+ ← Exp:C_44_01_ARM8.TXT+ Exp:C_38_18_BG.TXT+
28	Exp:C_42_14_ARM.TXT+ ← Exp:C_41_24_ARM8.TXT+ Exp:C_43_43_ARM8.TXT+
29	Exp:C_41_33_ARM8.TXT+ ← Exp:C_44_01_ARM8.TXT+ Exp:C_43_21_ARM8.TXT+

第 D章

付録

共通ルールを満たす遺伝子

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_45_ARM8.TXT+ Exp:C_43_21_ARM8.TXT+
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR181C RPS6B; 40S ribosomal protein S6e [SP:RS6_YEAST]
YDL191W RPL35A; 60S ribosomal protein L35e [SP:RL35_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL253W HXK2, HKB, HEX1; hexokinase II [EC:2.7.1.1] [SP:HXKB_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YJL138C TIF2, TIF41B; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAST]
YKR094C RPL40B; 60S ribosomal protein L40e [SP:RL40_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR075W RPL10; 60S ribosomal protein L10e [SP:RL10_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YPL079W RPL21B; 60S ribosomal protein L21e [SP:R21B_YEAST]
unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_24_ARM8.TXT+ Exp:C_38_18_BG.TXT+
known
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]
YDL130W RPP1B; 60S acidic ribosomal protein LP1 [SP:RLA3_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]

YDR226W ADK1; adenylate kinase [EC:2.7.4.3] [SP:KAD1_YEAST]
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGR214W RPS0A; 40S ribosomal protein SAe [SP:RS0A_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YKL006W RPL14A; 60S ribosomal protein L14e [SP:R14A_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YNL030W HHF2; histone H4 [SP:H4_YEAST]
YOR096W RPS7A; 40S ribosomal protein S7e [SP:RS7A_YEAST]
YPL143W RPL35A; 60S ribosomal protein L35Ae [SP:R33A_YEAST]
YPL081W RPS9A; 40S ribosomal protein S9e
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YLL044W unknown
YOR285W unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_19_ARM8.TXT+ Exp:C_42_44_ARM8.TXT+
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YDL191W RPL35A; 60S ribosomal protein L35e [SP:RL35_YEAST]
YDL136W RPL35B; 60S ribosomal protein L35e [SP:RL35_YEAST]
YER074W RPS24A; 40S ribosomal protein S24e [SP:RS24_YEAST]
YGL008C PMA1; H+-transporting ATPase [EC:3.6.1.35] [SP:PMA1_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YHR141C RPL44B; 60S ribosomal protein L44e [SP:RL44_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YJL177W RPL17B; 60S ribosomal protein L17e [SP:RL7B_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YLR075W RPL10; 60S ribosomal protein L10e [SP:RL10_YEAST]
YLR333C RPS25B; 40S ribosomal protein S25e [SP:RS25_YEAST]
YML063W RPS1B; 40S ribosomal protein S3Ae [SP:RS3B_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YNL067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YOL040C RPS15; 40S ribosomal protein S15e [SP:RS15_YEAST]
YOR063W RPL3; 60S ribosomal protein L3e [SP:RL3_YEAST]
YOR096W RPS7A; 40S ribosomal protein S7e [SP:RS7A_YEAST]
YPL143W RPL35A; 60S ribosomal protein L35Ae [SP:R33A_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YAR009C unknown
YCR013C unknown [SP:YCQ3_YEAST]
YLL044W unknown
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_43_21_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
known
YDL136W RPL35B; 60S ribosomal protein L35e [SP:RL35_YEAST]

```
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YIL018W RPL2B; 60S ribosomal proteins L8e [SP:RL6_YEAST]
YJL190C RPS15A; 40S ribosomal protein S15Ae [SP:RS22_YEAST]
YJL138C TIF2, TIF41B; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YJL052W TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
YPR043W RPL37A; 60S ribosomal protein L37Ae [SP:R37A_YEAST]
unknown
YAR009C unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YOR129C unknown

Exp:C_38_19_BG.TXT+ <- Exp:C_41_45_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
known
YDR050C TPI1; triosephosphate isomerase (TIM) [EC:5.3.1.1] [SP:TPIS_YEAST]
YDR064W RPS13; 40S ribosomal protein S13e [SP:RS13_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YFL039C ACT1; actin [SP:ACT_YEAST]
YGL147C RPL9A; 60S ribosomal protein L9e [SP:RL9A_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR441C RPS1A; 40S ribosomal protein S3Ae [SP:RS3A_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YPL220W RPL10A; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YPR080W TEF1; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YPR132W RPS23B; 40S ribosomal protein S23e [SP:RS28_YEAST]
unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YLR076C unknown
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_44_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
known
YCR031C RPS14A; 40S ribosomal protein S14e [SP:R141_YEAST]
YDL136W RPL35B; 60S ribosomal protein L35e [SP:RL35_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YEL009C GCN4; transcriptional activator of amino acid biosynthetic genes [SP:GCN4_YEAST]
YER117W RPL23B; 60S ribosomal protein L23e [SP:RL23_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
```

YJL052W TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YJR123W RPS5; 40S ribosomal protein S5e [SP:RS5_YEAST]
YLL045C RPL8B; 60S ribosomal protein L7Ae [SP:RL4B_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR340W RPP0; 60S acidic ribosomal protein LP0
YMR303C ADH2, ADR2; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH2_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YAR009C unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YOR129C unknown

Exp:C_41_22_ARM8.TXT+ <- Exp:C_41_34_ARM8.TXT+ Exp:C_41_40_ARM8.TXT+
known
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YBR191W RPL21A; 60S ribosomal protein L21e [SP:R21A_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGR148C RPL24B; 60S ribosomal protein L24e [SP:R24B_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YJL191W RPS14B; 40S ribosomal protein S14e [SP:R142_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR197W SIK1; similarity to microtubule binding proteins [SP:SIK1_YEAST]
YLR249W YEF3; elongation factor EF-3A [SP:EF3A_YEAST]
YLR264W RPS28B; 40S ribosomal protein S28e [SP:RS28_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL162W RPL44A; 60S ribosomal protein L44e [SP:RL44_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YOL039W RPP2A; 60S acidic ribosomal protein LP2 [SP:RLA2_YEAST]
YOR096W RPS7A; 40S ribosomal protein S7e [SP:RS7A_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
YPR035W GLN1; glutamine synthetase [EC:6.3.1.2] [SP:GLNA_YEAST]
unknown
YAR009C unknown
YLR339C unknown
YOL109W ZEO1; zeocin resistance

Exp:C_41_33_ARM8.TXT+ <- Exp:C_42_44_ARM8.TXT+ Exp:C_13_11_BG.TXT+
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YGL008C PMA1; H+-transporting ATPase [EC:3.6.1.35] [SP:PMA1_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YIL018W RPL2B; 60S ribosomal proteins L8e [SP:RL6_YEAST]
YLR333C RPS25B; 40S ribosomal protein S25e [SP:RS25_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YOR063W RPL3; 60S ribosomal protein L3e [SP:RL3_YEAST]
YOR096W RPS7A; 40S ribosomal protein S7e [SP:RS7A_YEAST]
unknown

YAR009C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_43_34_ARM8.TXT+ Exp:C_38_18_BG.TXT+
known
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR010W HHT1; histone H3 [SP:H3_YEAST]
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]
YBR189W RPS9B; 40S ribosomal protein S9e [SP:RS11_YEAST]
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YJR104C SOD1; superoxide dismutase (CU-ZN) [EC:1.15.1.1] [SP:SODC_YEAST]
YJR123W RPS5; 40S ribosomal protein S5e [SP:RS5_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR367W RPS15B; 40S ribosomal protein S15Ae
YML067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YML030W HHF2; histone H4 [SP:H4_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPL081W RPS9A; 40S ribosomal protein S9e
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YLL044W unknown

Exp:C_41_44_ARM8.TXT+ <- Exp:C_43_35_ARM8.TXT+ Exp:C_38_19_BG.TXT+
known
YBL087C RPL23A; 60S ribosomal protein L23e [SP:RL23_YEAST]
YCR012W PGK1; phosphoglycerate kinase [EC:2.7.2.3] [SP:PGK_YEAST]
YDR025W RPS11A; 40S ribosomal protein S11e [SP:RS11_YEAST]
YGL253W HXK2, HKB, HEX1; hexokinase II [EC:2.7.1.1] [SP:HXKB_YEAST]
YGL031C RPL24A; 60S ribosomal protein L24e [SP:R24A_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR214W RPS0A; 40S ribosomal protein SAe [SP:RSOA_YEAST]
YGR254W ENO1, ENOA, HSP48; enolase [EC:4.2.1.11] [SP:ENO1_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR134W PDC5; pyruvate decarboxylase isozyme 2 [EC:4.1.1.1] [SP:DCP2_YEAST]
YML178W RPS3; 40S ribosomal protein S3e [SP:RS3_YEAST]
YML162W RPL44A; 60S ribosomal protein L44e [SP:RL44_YEAST]
YML067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
unknown
YAR009C unknown
YDR154C unknown
YLR076C unknown

Exp:C_41_45_ARM8.TXT+ <- Exp:C_41_38_ARM8.TXT+ Exp:C_43_19_ARM8.TXT+
known
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDR050C TPI1; triosephosphate isomerase (TIM) [EC:5.3.1.1] [SP:TPIS_YEAST]
YDR385W EFT2; elongation factor EF-2 [EC:3.6.1.48] [SP:EF2_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]

YER117W RPL23B; 60S ribosomal protein L23e [SP:RL23_YEAST]
YGL123W RPS2; 40S ribosomal protein S2e [SP:RS4_YEAST]
YGR027C RPS25A; 40S ribosomal protein S25e [SP:RS25_YEAST]
YIL069C RPS24B; 40S ribosomal protein S24e [SP:RS24_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAST]
YWL209W SSB2; heat shock protein of HSP70 family, cytosolic [SP:HS76_YEAST]
YPR102C RPL11A; 60S ribosomal protein L11e [SP:RL11_YEAST]
unknown
YGL102C unknown [SP:YGK2_YEAST]
YLR076C unknown
YLR339C unknown

Exp:C_44_22_ARM8.TXT+ <- Exp:C_44_20_ARM8.TXT+ Exp:C_43_17_ARM8.TXT+
known
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDL136W RPL35B; 60S ribosomal protein L35e [SP:RL35_YEAST]
YDL081C RPP1A; 60S acidic ribosomal protein LP1 [SP:RLA1_YEAST]
YDR025W RPS11A; 40S ribosomal protein S11e [SP:RS11_YEAST]
YDR050C TPI1; triosephosphate isomerase (TIM) [EC:5.3.1.1] [SP:TPIS_YEAST]
YDR064W RPS13; 40S ribosomal protein S13e [SP:RS13_YEAST]
YGR209C TRX2; thioredoxin I (TR-I) [SP:TRX1_YEAST]
YHR010W RPL27A; 60S ribosomal protein L27e [SP:RL27_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR048W RPS0B; 40S ribosomal protein SAe [SP:RS0B_YEAST]
YWL031C HHT2, SIN2; histone H3 [SP:H3_YEAST]
YPR102C RPL11A; 60S ribosomal protein L11e [SP:RL11_YEAST]
unknown
YAR009C unknown
YIL152W unknown [SP:YIP2_YEAST]
YJL145W unknown [SP:YJO5_YEAST]

Exp:C_38_18_BG.TXT+ <- Exp:C_43_40_ARM8.TXT+ Exp:C_38_43_ARM8.TXT+
known
YAL038W CDC19, PYK1; pyruvate kinase 1 [EC:2.7.1.40] [SP:KPY1_YEAST]
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR010W HHT1; histone H3 [SP:H3_YEAST]
YCR012W PGK1; phosphoglycerate kinase [EC:2.7.2.3] [SP:PGK_YEAST]
YDR382W RPP2B; 60S acidic ribosomal protein LP2 [SP:RLA4_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YKL216W URA1; dihydroorotate oxidase [EC:1.3.3.1] [SP:PYRD_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR048W RPS0B; 40S ribosomal protein SAe [SP:RS0B_YEAST]
YLR167W RPS31, UBI3; 40S ribosomal protein S27Ae
YLR367W RPS15B; 40S ribosomal protein S15Ae
YPL220W RPL10A; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YLL044W unknown

Exp:C_41_24_ARM8.TXT+ <- Exp:C_43_45_ARM8.TXT+ Exp:C_41_18_ARM8.TXT+
known
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]

```
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR226W ADK1; adenylate kinase [EC:2.7.4.3] [SP:KAD1_YEAST]
YDR447C RPS17B; 40S ribosomal protein S17e
YEL054C RPL12A, RPL15B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL030W RPL30; 60S ribosomal protein L30e [SP:RL30_YEAST]
YGL008C PMA1; H+-transporting ATPase [EC:3.6.1.35] [SP:PMA1_YEAST]
YGR214W RPS0A; 40S ribosomal protein SAe [SP:RS0A_YEAST]
YHL033C RPL8A; 60S ribosomal protein L7Ae [SP:RL4A_YEAST]
YKL180W RPL17A; 60S ribosomal protein L17e [SP:RL7A_YEAST]
YLL045C RPL8B; 60S ribosomal protein L7Ae [SP:RL4B_YEAST]
YLL039C UBI4; ubiquitin
YLR134W PDC5; pyruvate decarboxylase isozyme 2 [EC:4.1.1.1] [SP:DCP2_YEAST]
YPL081W RPS9A; 40S ribosomal protein S9e
unknown
YGL102C unknown [SP:YGK2_YEAST]
YLR076C unknown
YMR027W HRT2; high level expression reduced Ty3 transposition [SP:YMR7_YEAST]

Exp:C_41_44_ARM8.TXT+ <- Exp:C_44_17_ARM8.TXT+ Exp:C_20_21_BG.TXT+
known
YBR196C PGI1; glucose-6-phosphate isomerase [EC:5.3.1.9] [SP:G6PI_YEAST]
YDL130W RPP1B; 60S acidic ribosomal protein LP1 [SP:RLA3_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR025W RPS11A; 40S ribosomal protein S11e [SP:RS11_YEAST]
YDR155C CPH1, CPR1, CYP1, SCC1; cyclophilin peptidyl-prolyl cis-trans isomerase [EC:5.2.1.8] [SP:CYPH_YEAST]
YEL054C RPL12A, RPL15B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL103W RPL28; 60S ribosomal protein L27Ae [SP:R27A_YEAST]
YGR027C RPS25A; 40S ribosomal protein S25e [SP:RS25_YEAST]
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR134W PDC5; pyruvate decarboxylase isozyme 2 [EC:4.1.1.1] [SP:DCP2_YEAST]
YML067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPR102C RPL11A; 60S ribosomal protein L11e [SP:RL11_YEAST]
unknown
YAR009C unknown
YKL153W unknown [SP:YKP3_YEAST]
YLR076C unknown

Exp:C_44_17_ARM8.TXT+ <- Exp:C_43_41_ARM8.TXT+ Exp:C_43_23_ARM8.TXT+
known
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YJL190C RPS15A; 40S ribosomal protein S15Ae [SP:RS22_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR134W PDC5; pyruvate decarboxylase isozyme 2 [EC:4.1.1.1] [SP:DCP2_YEAST]
YML063W RPS1B; 40S ribosomal protein S3Ae [SP:RS3B_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPL220W RPL10A; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
```

```
YPR102C RPL11A; 60S ribosomal protein L11e [SP:RL11_YEAST]
unknown
YCR013C unknown [SP:YCQ3_YEAST]
YGL102C unknown [SP:YGK2_YEAST]
YLR076C unknown

Exp:C_44_18_ARM8.TXT+ <- Exp:C_41_24_ARM8.TXT+ Exp:C_43_35_ARM8.TXT+
known
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBL005W PDR3; pleiotropic drug resistance regulatory protein 3 [SP:PDR3_YEAST]
YBR009C HHF1; histone H4 [SP:H4_YEAST]
YBR048W RPS11B; 40S ribosomal protein S11e [SP:RS11_YEAST]
YCR012W PGK1; phosphoglycerate kinase [EC:2.7.2.3] [SP:PGK_YEAST]
YDR447C RPS17B; 40S ribosomal protein S17e
YGL103W RPL28; 60S ribosomal protein L27Ae [SP:R27A_YEAST]
YGL030W RPL30; 60S ribosomal protein L30e [SP:RL30_YEAST]
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YGR209C TRX2; thioredoxin I (TR-I) [SP:TRX1_YEAST]
YJL177W RPL17B; 60S ribosomal protein L17e [SP:RL7B_YEAST]
YJL052W TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YKL216W URA1; dihydroorotate oxidase [EC:1.3.3.1] [SP:PYRD_YEAST]
YKL180W RPL17A; 60S ribosomal protein L17e [SP:RL7A_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YWL030W HHF2; histone H4 [SP:H4_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YAR009C unknown
YDR154C unknown
YLR076C unknown

Exp:C_41_24_ARM8.TXT+ <- Exp:C_42_12_ARM.TXT+ Exp:C_43_34_ARM8.TXT+
known
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YDL081C RPP1A; 60S acidic ribosomal protein LP1 [SP:RLA1_YEAST]
YDR050C TPI1; triosephosphate isomerase (TIM) [EC:5.3.1.1] [SP:TPIS_YEAST]
YDR385W EFT2; elongation factor EF-2 [EC:3.6.1.48] [SP:EF2_YEAST]
YER131W RPS26B; 40S ribosomal protein S26e [SP:R26B_YEAST]
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGL123W RPS2; 40S ribosomal protein S2e [SP:RS4_YEAST]
YGL030W RPL30; 60S ribosomal protein L30e [SP:RL30_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YKL180W RPL17A; 60S ribosomal protein L17e [SP:RL7A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR344W RPL26A; 60S ribosomal protein L26e
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
YPL081W RPS9A; 40S ribosomal protein S9e
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOL109W ZEO1; zeocin resistance

Exp:C_41_47_ARM.TXT+ <- Exp:C_42_12_ARM.TXT+ Exp:C_38_44_ARM8.TXT+
known
```

YAL038W CDC19, PYK1; pyruvate kinase 1 [EC:2.7.1.40] [SP:KPY1_YEAST]
YDR050C TPI1; triosephosphate isomerase (TIM) [EC:5.3.1.1] [SP:TPIS_YEAST]
YDR345C HXT3; low-affinity hexose transporter [SP:HXT3_YEAST]
YDR385W EFT2; elongation factor EF-2 [EC:3.6.1.48] [SP:EF2_YEAST]
YER102W RPS8B; 40S ribosomal protein S8e [SP:RS8_YEAST]
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGL147C RPL9A; 60S ribosomal protein L9e [SP:RL9A_YEAST]
YJL190C RPS15A; 40S ribosomal protein S15Ae [SP:RS22_YEAST]
YJR123W RPS5; 40S ribosomal protein S5e [SP:RS5_YEAST]
YLL045C RPL8B; 60S ribosomal protein L7Ae [SP:RL4B_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR441C RPS1A; 40S ribosomal protein S3Ae [SP:RS3A_YEAST]
YMR303C ADH2, ADR2; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH2_YEAST]
YOR167C RPS28A; 40S ribosomal protein S28e [SP:RS28_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YAR009C unknown
YDR154C unknown
YLR076C unknown

Exp:C_41_19_ARM8.TXT+ <- Exp:C_41_44_ARM8.TXT+ Exp:C_43_42_ARM8.TXT+
known
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YGR214W RPS0A; 40S ribosomal protein SAe [SP:RS0A_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YHR203C RPS4B; 40S ribosomal protein S4e [SP:RS4_YEAST]
YKL180W RPL17A; 60S ribosomal protein L17e [SP:RL7A_YEAST]
YLL045C RPL8B; 60S ribosomal protein L7Ae [SP:RL4B_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR249W YEF3; elongation factor EF-3A [SP:EF3A_YEAST]
YLR333C RPS25B; 40S ribosomal protein S25e [SP:RS25_YEAST]
YLR340W RPP0; 60S acidic ribosomal protein LP0
YML069C RPL16B; 60S ribosomal protein L13Ae [SP:R16B_YEAST]
YML067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
unknown
YAR009C unknown
YLL044W unknown
YOL109W ZEO1; zeocin resistance

Exp:C_41_26_ARM8.TXT+ <- Exp:C_44_01_ARM8.TXT+ Exp:C_38_18_BG.TXT+
known
YBR010W HHT1; histone H3 [SP:H3_YEAST]
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YCR012W PGK1; phosphoglycerate kinase [EC:2.7.2.3] [SP:PGK_YEAST]
YDL081C RPP1A; 60S acidic ribosomal protein LP1 [SP:RLA1_YEAST]
YGL030W RPL30; 60S ribosomal protein L30e [SP:RL30_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YJL159W HSP150; heat shock protein, secretory glycoprotein [SP:CCW7_YEAST]
YKL216W URA1; dihydroorotate oxidase [EC:1.3.3.1] [SP:PYRD_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]

YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YMR303C ADH2, ADR2; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH2_YEAST]
YNL162W RPL44A; 60S ribosomal protein L44e [SP:RL44_YEAST]
YNL030W HHF2; histone H4 [SP:H4_YEAST]
unknown
YLL044W unknown
YLR162W unknown
YOR285W unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_43_34_ARM8.TXT+ Exp:C_38_44_ARM8.TXT+
known
YEL009C GCN4; transcriptional activator of amino acid biosynthetic genes [SP:GCN4_YEAST]
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGL008C PMA1; H+-transporting ATPase [EC:3.6.1.35] [SP:PMA1_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YIL018W RPL2B; 60S ribosomal proteins L8e [SP:RL6_YEAST]
YJL052W TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YJR123W RPS5; 40S ribosomal protein S5e [SP:RS5_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YNL301C RPL18B; 60S ribosomal protein L18e [SP:RL18_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_45_ARM8.TXT+ Exp:C_38_18_BG.TXT+
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YBL027W RPL19B; 60S ribosomal protein L19e [SP:RL19_YEAST]
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAST]
YJR104C SOD1; superoxide dismutase (CU-ZN) [EC:1.15.1.1] [SP:SODC_YEAST]
YKL006W RPL14A; 60S ribosomal protein L14e [SP:R14A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR075W RPL10; 60S ribosomal protein L10e [SP:RL10_YEAST]
YLR367W RPS15B; 40S ribosomal protein S15Ae
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YNL067W RPL9B; 60S ribosomal protein L9e [SP:RL9B_YEAST]
YNL030W HHF2; histone H4 [SP:H4_YEAST]
YPL143W RPL35A; 60S ribosomal protein L35Ae [SP:R33A_YEAST]
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOR285W unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_41_23_ARM8.TXT+ Exp:C_38_18_BG.TXT+

known
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]
YBR189W RPS9B; 40S ribosomal protein S9e [SP:RS11_YEAST]
YBR191W RPL21A; 60S ribosomal protein L21e [SP:R21A_YEAST]
YCR031C RPS14A; 40S ribosomal protein S14e [SP:R141_YEAST]
YGR214W RPS0A; 40S ribosomal protein SAe [SP:RS0A_YEAST]
YHL001W RPL14B; 60S ribosomal protein L14e [SP:R14B_YEAST]
YJL159W HSP150; heat shock protein, secretory glycoprotein [SP:CCW7_YEAST]
YJR009C TDH2, GPD2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P2_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YJR123W RPS5; 40S ribosomal protein S5e [SP:RS5_YEAST]
YKL006W RPL14A; 60S ribosomal protein L14e [SP:R14A_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR029C RPL15A; 60S ribosomal protein L15e [SP:R15A_YEAST]
YLR048W RPS0B; 40S ribosomal protein SAe [SP:RS0B_YEAST]
YLR075W RPL10; 60S ribosomal protein L10e [SP:RL10_YEAST]
YLR367W RPS15B; 40S ribosomal protein S15Ae
YMR303C ADH2, ADR2; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH2_YEAST]
YNL030W HHF2; histone H4 [SP:H4_YEAST]
YOL086C ADH1, ADC1; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH1_YEAST]
YPL143W RPL35A; 60S ribosomal protein L35Ae [SP:R33A_YEAST]
YPL081W RPS9A; 40S ribosomal protein S9e
unknown
YDR154C unknown
YGL102C unknown [SP:YGK2_YEAST]
YOR285W unknown

Exp:C_44_09_ARM8.TXT+ <- Exp:C_44_15_ARM8.TXT+ Exp:C_44_03_ARM8.TXT+
known
YBR010W HHT1; histone H3 [SP:H3_YEAST]
YCR012W PGK1; phosphoglycerate kinase [EC:2.7.2.3] [SP:PGK_YEAST]
YDR064W RPS13; 40S ribosomal protein S13e [SP:RS13_YEAST]
YDR226W ADK1; adenylate kinase [EC:2.7.4.3] [SP:KAD1_YEAST]
YDR447C RPS17B; 40S ribosomal protein S17e
YIL052C RPL34B; 60S ribosomal protein L34e [SP:R34B_YEAST]
YIR032C DAL3; ureidoglycolate hydrolase [EC:3.5.3.19] [SP:DAL3_YEAST]
YJL159W HSP150; heat shock protein, secretory glycoprotein [SP:CCW7_YEAST]
YJL124C LSM1; suppressor of PAB1 [SP:YJM4_YEAST]
YJL117W PHO86; inorganic phosphate transporter [SP:PH86_YEAST]
YKL152C GPM1; phosphoglycerate mutase [EC:5.4.2.1] [SP:PMG1_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR167W RPS31, UBI3; 40S ribosomal protein S27Ae
unknown
YIL059C unknown [SP:YIF9_YEAST]
YJL104W unknown [SP:YJK4_YEAST]
YLR076C unknown

Exp:C_44_07_ARM8.TXT+ <- Exp:C_44_21_ARM8.TXT+ Exp:C_43_19_ARM8.TXT+
known
YAL005C SSA1; heat shock protein of HSP70 family, cytosolic [SP:HS71_YEAST]
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDR155C CPH1, CPR1, CYP1, SCC1; cyclophilin peptidyl-prolyl cis-trans isomerase [EC:5.2.1.8] [SP:CYPH_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YHR174W ENO2, ENOB; enolase [EC:4.2.1.11] [SP:ENO2_YEAST]

YLR058C SHM2; serine hydroxymethyltransferase, cytosolic (glycine hydroxymethyltransferase) [EC:2.1.2.1] [SP:GLYC_YEAST]
YML028W TSA1; TSA TSA1; thiol-specific antioxidant [SP:TSA1_YEAST]
YNL209W SSB2; heat shock protein of HSP70 family, cytosolic [SP:HS76_YEAST]
YOR369C RPS12; 40S ribosomal protein S12e [SP:RS12_YEAST]
YPL131W RPL5; 60S ribosomal protein L5e [SP:RL5_YEAST]
unknown
YGL102C unknown [SP:YGK2_YEAST]
YIL152W unknown [SP:YIP2_YEAST]
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_44_01_ARM8.TXT+ Exp:C_38_18_BG.TXT+
known
YBR010W HHT1; histone H3 [SP:H3_YEAST]
YBR031W RPL4A; 60S ribosomal protein L4e [SP:RL4A_YEAST]
YBR118W TEF2; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YHL001W RPL14B; 60S ribosomal protein L14e [SP:R14B_YEAST]
YJL159W HSP150; heat shock protein, secretory glycoprotein [SP:CCW7_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YJR104C SOD1; superoxide dismutase (CU-ZN) [EC:1.15.1.1] [SP:SODC_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YLR367W RPS15B; 40S ribosomal protein S15Ae
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
YMR303C ADH2, ADR2; alcohol dehydrogenase [EC:1.1.1.1] [SP:ADH2_YEAST]
YNL030W HHF2; histone H4 [SP:H4_YEAST]
unknown
YDR154C unknown
YLL044W unknown
YOR285W unknown

Exp:C_42_14_ARM.TXT+ <- Exp:C_41_24_ARM8.TXT+ Exp:C_43_43_ARM8.TXT+
known
YBR009C HHF1; histone H4 [SP:H4_YEAST]
YDL081C RPP1A; 60S acidic ribosomal protein LP1 [SP:RLA1_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR447C RPS17B; 40S ribosomal protein S17e
YFL045C SEC53; phosphomannomutase [EC:5.4.2.8] [SP:PMM_YEAST]
YGL030W RPL30; 60S ribosomal protein L30e [SP:RL30_YEAST]
YGR085C RPL11B; 60S ribosomal protein L11e [SP:RL11_YEAST]
YJR047C ANB1, HYP1, TIF51B; initiation factor eIF-5A-1 [SP:IF51_YEAST]
YKR059W TIF1, TIF41A; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YLL045C RPL8B; 60S ribosomal protein L7Ae [SP:RL4B_YEAST]
YLR044C PDC1; pyruvate decarboxylase isozyme 1 [EC:4.1.1.1] [SP:DCP1_YEAST]
YLR134W PDC5; pyruvate decarboxylase isozyme 2 [EC:4.1.1.1] [SP:DCP2_YEAST]
YLR441C RPS1A; 40S ribosomal protein S3Ae [SP:RS3A_YEAST]
YPR080W TEF1; elongation factor EF-1 alpha subunit [EC:3.6.1.48] [SP:EF1A_YEAST]
unknown
YLL044W unknown
YOL109W ZEO1; zeocin resistance
YOR129C unknown

Exp:C_41_33_ARM8.TXT+ <- Exp:C_44_01_ARM8.TXT+ Exp:C_43_21_ARM8.TXT+
known

YBR010W HHT1; histone H3 [SP:H3_YEAST]
YBR181C RPS6B; 40S ribosomal protein S6e [SP:RS6_YEAST]
YDR012W RPL4B; 60S ribosomal protein L4e [SP:RL4B_YEAST]
YDR418W RPL12B; 60S ribosomal protein L12e [SP:RL12_YEAST]
YGL135W RPL1B; 60S ribosomal protein L10Ae [SP:R10A_YEAST]
YGR118W RPS23A; 40S ribosomal protein S23e [SP:RS28_YEAST]
YGR192C TDH3, GPD3; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P3_YEAST]
YHR141C RPL44B; 60S ribosomal protein L44e [SP:RL44_YEAST]
YIL018W RPL2B; 60S ribosomal proteins L8e [SP:RL6_YEAST]
YJL177W RPL17B; 60S ribosomal protein L17e [SP:RL7B_YEAST]
YJL138C TIF2, TIF41B; eukaryotic initiation factor eIF-4A [SP:IF4A_YEAST]
YJL052W TDH1, GPD1, SSS2; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12] [SP:G3P1_YEAST]
YKR094C RPL40B; 60S ribosomal protein L40e [SP:RL40_YEAST]
YLL024C SSA2; heat shock protein of HSP70 family, cytosolic [SP:HS72_YEAST]
YMR121C RPL15B; 60S ribosomal protein L15e [SP:R15B_YEAST]
unknown
YAR009C unknown
YCR013C unknown [SP:YCQ3_YEAST]
YDR154C unknown