

(19) 世界知的所有権機関
国際事務局(43) 国際公開日
2007年2月8日 (08.02.2007)

PCT

(10) 国際公開番号
WO 2007/015459 A1

- (51) 国際特許分類:
C12Q 1/68 (2006.01) C12N 15/09 (2006.01)
- (21) 国際出願番号: PCT/JP2006/315143
- (22) 国際出願日: 2006年7月31日 (31.07.2006)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願2005-222995 2005年8月1日 (01.08.2005) JP
- (71) 出願人 (米国を除く全ての指定国について): 国立大学法人大阪大学 (OSAKA UNIVERSITY) [JP/JP]; 〒5650871 大阪府吹田市山田丘1番1号 Osaka (JP).
- (72) 発明者: および
- (75) 発明者/出願人 (米国についてのみ): 竹政 伊知朗 (TAKEMASA, Ichiro) [JP/JP]; 〒5650871 大阪府吹田市山田丘1番1号 国立大学法人大阪大学内 Osaka (JP). 田崎 武信 (TASAKI, Takenobu) [JP/JP]; 〒5530002 大阪府大阪市福島区鷺洲5丁目12番4号 塩野義製薬株式会社内 Osaka (JP). 園田 光 (SONODA, Hikaru) [JP/JP]; 〒5530002 大阪府大阪市福島区鷺洲5丁目12番4号 塩野義製薬株式会社内 Osaka (JP). 樋口 浩文 (HIGUCHI, Hirofumi) [JP/JP]; 〒8691298 熊本県

菊池市旭志川辺1314-1 財団法人化学及血清療法研究所菊池研究所内 Kumamoto (JP). 松原 謙一 (MATSUBARA, Kenichi) [JP/JP]; 〒5650821 大阪府吹田市山田東3丁目18番1-804号 Osaka (JP).

(74) 代理人: 河宮 治, 外 (KAWAMIYA, Osamu et al.); 〒5400001 大阪府大阪市中央区城見1丁目3番7号 IMPビル青山特許事務所 Osaka (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR),

[続葉有]

(54) Title: GENE SET FOR USE IN PREDICTION OF OCCURRENCE OF LYMPH NODE METASTASIS OF COLORECTAL CANCER

(54) 発明の名称: 大腸癌リンパ節転移の有無を予測するための遺伝子セット

(57) Abstract: Disclosed is a method for prediction of the presence or absence of lymph node metastasis of colorectal cancer. Also disclosed is a set of genes which can be used in the method. A method for selecting a set of genes for use in the prediction of the presence or absence of lymph node metastasis of colorectal cancer, the method comprising the following steps (1) to (4): (1) analyzing the information about gene expression in a primary lesion of colorectal cancer in a patient who has been determined to have lymph node metastasis of colorectal cancer by a histopathological test, by at least four analysis methods involving at least one supervised learning analysis method, thereby selecting a group of genes which serve to determine the presence or absence of lymph node metastasis at a correct classification rate of 75% or higher for each of the analysis methods; (2) selecting a common gene which is selected from the group of genes in all of the analysis methods of step (1); (3) analyzing the information about gene expression above to assign the presence or absence of lymph node metastasis to any combination of two or more genes and then selecting a combination(s) of genes showing an interaction from the combinations; and (4) performing a variable selection in a logistic regression model to give an answer regarding the presence or absence of lymph node metastasis using the common gene and the combination(s) of genes as the explaining variables; a set of genes for use in the prediction of the presence or absence of lymph node metastasis of colorectal cancer, which is selected by the method; and a method for prediction of the presence or absence of lymph node metastasis of colorectal cancer by using the set of genes.

(57) 要約: 大腸癌のリンパ節転移の有無を予測するための方法及び当該方法に利用できる遺伝子セットを提供する。下記(1)～(4)の工程を含む、大腸癌リンパ節転移の有無を予測するための遺伝子セットの選択方法: (1) 組織病理学的判定によりリンパ節転移の有無が明らかにされた患者の大腸癌原発巣組織における遺伝子発現情報を、教師あり学習解析方法を少なくとも一つ含む、4以上の解析方法で解析することにより、リンパ節転移の有無を正分類率75%以上で分類できる遺伝子群をそれぞれの解析方法において選定する工程、(2) (1)で用いたそれぞれの解析方法で選定された遺伝子群から、何れの解析方法でも共通して選定された共通遺伝子を選択する工程、(3) 前記遺伝子発現情報を解析することにより、任意の2以上の遺伝子の組合せの中から、リンパ節転移の有無の分類を指示し、交互作用を示す遺伝子の組合せを選択する工程、及び(4) 前記共通遺伝子及び前記遺伝子の組合せを説明変数として、リンパ節転移の有無を応答としたロジスティック回帰モデルにおける変数選択を行う工程; 該選択方法により選択される、大腸癌リンパ節転移の有無を予測するための遺伝子セット; 及び該遺伝子セットを用いることを特徴とする大腸癌リンパ節転移の有無を予測するための方法。

WO 2007/015459 A1



OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML,
MR, NE, SN, TD, TG).

2文字コード及び他の略語については、定期発行される
各PCTガゼットの巻頭に掲載されている「コードと略語
のガイダンスノート」を参照。

添付公開書類:

— 国際調査報告書

明 細 書

大腸癌リンパ節転移の有無を予測するための遺伝子セット

技術分野

[0001] 本発明は、大腸癌のリンパ節転移の有無を予測するのに有用な遺伝子群、及びそれらの遺伝子発現情報の活用法に関する。

背景技術

[0002] 大腸癌は先進国では特に発生率が高く、本邦でも年々増加の一途を辿っており、癌関連死の主要な原因のひとつとなっている。統計学的な報告(例えば、非特許文献1参照)によれば、大腸癌患者の37%は転移がない主病巣に局限した癌であり、同じく37%が所属リンパ節にのみ転移を認める局在的な癌であり、残りは遠隔転移を伴っているものなどであることが分かっている。

[0003] 現在、臨床における大腸癌の悪性度分類として最も一般的に利用されているDukes分類においては、癌の大腸壁深達度と所属リンパ節への転移の程度などの病理学的な事項を指標としており、本分類と予後の相関に関しては疑う余地のないところとなっている。しかしながら、上記分類結果を大きく左右するリンパ節転移の有無の判定に関しては、切除された数多くのリンパ節組織のうちの一部を用いて作成された標本を顕微鏡下で観察するという古典的な組織病理学的手法に頼っているのが現状である。このような手法によりリンパ節転移陰性と判定された患者の20～40%には後に転移が発見されるとの報告(例えば、非特許文献2参照)が示すように、従来のリンパ節転移判定法は必ずしも精度が十分とはいえなかった。

[0004] 大腸癌は、多段階発癌の構造など分子生物学的な研究がもっともよく進んでいる癌の一つで、これまでAPC、K-ras、p53、DCCなどの個々の遺伝子についての報告が多数みられる。しかし、これらの遺伝子のいずれかに注目するだけでは、大腸癌の個性を表現するには不十分であるため、近年はDNAマイクロアレイなどを用いることにより一度に極めて多数の遺伝子の発現情報を得ることにより有用な新規知見を得る試みがなされ始めている。

[0005] Alizadehらは、びまん性大細胞型B細胞リンパ腫患者の末梢血から分取したBリン

パ球を試料としてDNAマイクロアレイによる測定を行い、得られた遺伝子発現データの階層的クラスタリングを行うことにより、同病患者の末梢血Bリンパ球には、リンパ組織の胚中心に存在するB細胞に類似した遺伝子発現パターンを示す場合と、in vitroで活性化したB細胞に類似した遺伝子発現パターンを示す場合の2種類があることを見出した(非特許文献3)。両者の生存率をKaplan-Meierプロットで調べた結果、後者の発現パターンを示すB細胞を持つ患者は、前者の発現パターンを示すB細胞を持つ患者と比べて予後が悪いことが明らかとなった。加えて、従来からの病理学的診断に基づく予後予測に従うよりも、著者らの行った遺伝子発現情報のクラスタリングで得られた結果の方が予後との相関性が高かった。Alizadehらの研究結果は、遺伝子発現情報から臨床的に利用可能な有用な法則性を導き出したという点で意義のあるものといえる。しかし、その法則が全く新たな臨床例についても適用できるかどうかについての検証はなされておらず、この論文の範囲でのみ成立する結果である可能性は否定できない。

[0006] Khanらは、組織学的には区別が難しい小円形青色細胞腫に属する4種類の癌が、人工ニューラルネットワークを利用した遺伝子発現情報の解析により正確に区別されることを報告した(非特許文献4)。この報告の中では、全体のデータから無作為に抜き出した一部のデータを用いて導き出した人工ニューラルネットワークモデルに対して、テストサンプルのデータを入力した場合にも、正確な判定結果が得られることが検証されている。したがって、ここで導き出された人工ニューラルネットワークモデルは、この論文内のデータの範囲に限定されるものではなく、小円形青色細胞腫に属する4種類の癌を区別するために一般的に適用可能なものであることが示唆される。しかしながら、人工ニューラルネットワークモデルで得られる判定結果は、数学的な根拠を明確に説明できないという点で一般には受け入れられにくい。

[0007] 大腸癌の肝転移に関わる分子標的を同定することを目的としてDNAマイクロアレイを用いて行われた最近の研究例としては、柳川らの報告(非特許文献5)がある。著者らは、公共の遺伝子データベースに登録されているヒトcDNAの塩基配列に基づいて設計したオリゴDNAをプライマーとして用い、ヒトのcDNAを鋳型としてPCRを行い、9,121種類の増幅cDNA断片を得た。次いで、これらのcDNA断片をプロー

ブとしてプリントしたDNAマイクロアレイを使って、10症例の大腸癌患者より分離した大腸癌原発巣及び大腸癌肝転移巣の遺伝子発現プロファイルを調べた。その結果、原発巣に対して肝転移巣で発現が上昇している40種類の遺伝子と、原発巣に対して肝転移巣で発現が低下している7種類の遺伝子を明らかにし、大腸癌の肝転移に関わる可能性がある候補遺伝子セットを同定した。

[0008] 大腸癌の肝転移に関与する遺伝子セットについては、DNAマイクロアレイ法により大腸癌原発巣組織に特異的に発現した遺伝子群の発現情報を遺伝子判別分析手法に基づく統計解析処理することにより、大腸癌の肝転移の予測に有効な遺伝子セットを同定する方法、当該方法によって同定された遺伝子セット及び大腸癌原発巣組織における当該遺伝子セットの発現情報を用いて大腸癌の肝転移を予測する方法が知られている(特許文献1)。当該遺伝子セット及び方法は、大腸癌の異時性肝転移の予測に有用な情報を提供するものであり、大腸癌で特異的に発現している重要な遺伝子を同定するための材料として好ましいものではあるが、大腸癌のリンパ節転移は肝転移とは病理学的にみて全く病態が異なるため、これら大腸癌の肝転移用の遺伝子セット及び方法をそのまま大腸癌のリンパ節転移に応用できるわけでは決していない。

[0009] また、大腸癌原発巣組織、大腸癌肝転移巣組織及び正常大腸粘膜組織を材料として作製したcDNAライブラリーから選択したプローブを用いてオリジナルのDNAマイクロアレイを作製し、それを用いて大腸癌組織における遺伝子発現解析を行うことにより、大腸癌の発育・進展に関連すると考えられる候補遺伝子の同定が可能であることも示されている(非特許文献6)。

[0010] 一方、大腸癌のリンパ節転移に関しては、上記のようにリンパ節転移の有無の判定が、切除された数多くのリンパ節組織のうちの一部を用いて作成された標本を顕微鏡下で観察するという古典的な組織病理学的手法に頼っているのが現状であり、このようなリンパ節転移判定法は精度が必ずしも十分とはいえない。また、大腸癌原発巣除去手術後に行われる術後補助療法により、リンパ節転移のあった患者の予後を改善できることも知られているが、術後補助療法は食欲不振・上腹部不快感・嘔気などの副作用を伴うこともあり、Quality of life (QOL) や医療費の観点からみて、患者個人の

状態と病勢を考慮して必要・不必要を判断する必要がある。従って、より精度の高いリンパ節転移判定法が見出されれば、術後補助療法の選択に際して意志決定のための有用な指標として利用可能となり、最終的には適切な治療が受けられることにより患者の利益につながると考えられる。

[0011] 特許文献1:特開2004-33082

非特許文献1:Troisi R.J.ら、1999, Cancer, vol. 85, p. 1670-1676

非特許文献2:Cohen A.M.ら、1997, Curr Probl Surg., vol. 34, p. 601-676

非特許文献3:Alizadehら、2000, Nature, vol. 403, p. 503-511

非特許文献4:Khanら、2001, Nature Medicine, vol. 7, p. 673-679

非特許文献5:柳川ら、2001, Neoplasia, vol. 3, No. 5, p.395-401

非特許文献6:竹政ら、2001, Biochem. Biophys. Res. Commun., vol. 285, p. 1244-1249

発明の開示

発明が解決しようとする課題

[0012] 上記のように、大腸癌のリンパ節転移の有無を判定する従来の方法は、大腸癌周辺の複数のリンパ節を切除し、これを顕微鏡下で観察するもので、判定結果の精度に問題があった。

[0013] したがって、本発明は、この点を改善するために、大腸癌原発巣組織の遺伝子発現プロファイルを調べることにより、大腸癌のリンパ節転移の有無を予測する方法を提供することを目的とする。本発明はまた、リンパ節への癌細胞の転移の有無を予測することを可能とならしめるために、大腸癌のリンパ節転移判定に利用可能な遺伝子セット及びそれらの発現情報に基づいてリンパ節転移の有無を判定するために利用可能な判別式を提供することを目的とする。

課題を解決するための手段

[0014] 本発明者らは、上記の目的を達成するために鋭意研究を重ねた結果、大腸癌原発巣組織、大腸癌肝転移巣組織及び正常大腸粘膜組織を材料として作製したcDNAライブラリーから選択したプローブを用いてオリジナルのDNAマイクロアレイを作製し、当該DNAマイクロアレイを用いて得た大腸癌原発巣の遺伝子発現解析データの

統計解析を通じて、リンパ節転移の有無を予測するのに利用可能な遺伝子セット、及びそれらの発現量に基づいて実際にリンパ節転移の有無を予測するために利用する判別式を見出すことに成功し、本発明を完成するに至った。

[0015] すなわち、本発明は、以下の大腸癌のリンパ節転移の有無を予測するための遺伝子セットの選択方法を提供する。

1. 下記(1)～(4)の工程を含む、大腸癌リンパ節転移の有無を予測するための遺伝子セットの選択方法：

(1)組織病理学的判定によりリンパ節転移の有無が明らかにされた患者の大腸癌原発巣組織における遺伝子発現情報を、教師あり学習解析方法を少なくとも一つ含む、4以上の解析方法で解析することにより、リンパ節転移の有無を正分類率75%以上で分類できる遺伝子群をそれぞれの解析方法において選定する工程、

(2) (1)で用いたそれぞれの解析方法で選定された遺伝子群から、何れの解析方法でも共通して選定された共通遺伝子を選択する工程、

(3)前記遺伝子発現情報を解析することにより、任意の2以上の遺伝子の組合せの中から、リンパ節転移の有無の分類を指示し、交互作用を示す遺伝子の組合わせを選択する工程、及び

(4)前記共通遺伝子及び前記遺伝子の組合わせを説明変数として、リンパ節転移の有無を応答としたロジスティック回帰モデルにおける変数選択を行う工程；

2. (1)の解析方法が、(a)Support Vector Machine、(b)Principal Component Analysis Artificial Neural Networkの拡張法、(c)Hierarchical Cluster AnalysisとStepwise Logistic Discriminationの組合せ及び(d)Classification And Regression TreeとLogistic Discriminationの組合せよりなる群から選択されるものを少なくとも一つ含むものである、上記1. に記載の方法；

3. (3)の解析方法が、(d)Classification And Regression TreeとLogistic Discriminationの組合せである、上記1. または2. に記載の方法；

4. (4)の変数選択の方法が、ステップワイズの変数選択法である、上記1. ないし3. のいずれかに記載の方法。

[0016] 本発明はまた、以下の大腸癌リンパ節転移の有無を予測するための遺伝子セットを

提供する。

5. 上記1. ないし4. のいずれかの方法により選択される、大腸癌リンパ節転移の有無を予測するための遺伝子セット；

6. 少なくともNM_003404 (G1592)、NM_002128 (G2645)、NM_052868 (G3031)、NM_005034 (G3177)、NM_001540 (G3753)、NM_005722 (G3826)、及びNM_015315 (G4370)のデータベースのアクセス番号(シリアル番号)で表される遺伝子を含む、上記5. に記載の遺伝子セット。

[0017] 本発明はさらに、以下の、上記選択された遺伝子セットを用いた大腸癌のリンパ節転移の有無を予測する方法をも提供する。

7. 上記5. または6. の何れかに記載の遺伝子セットを用いることを特徴とする大腸癌リンパ節転移の有無を予測するための方法；

8. 下記の判別式を用いることを特徴とする上記7. 記載の方法：

$$\begin{aligned}
 D = & 0.2307 - 2.7132 \times \text{NM\_003404 (G1592) の発現量} \\
 & + 8.9509 \times \text{NM\_052868 (G3031) の発現量} \\
 & + 8.7975 \times \text{NM\_005722 (G3826) の発現量} \\
 & - 2.3098 \times \text{NM\_015315 (G4370) の発現量} \\
 & + 3.5126 \times \text{NM\_002128 (G2645) の発現量} \times \text{NM\_005034 (G3177) の発現量} \\
 & - 8.8226 \times \text{NM\_001540 (G3753) の発現量} \times \text{NM\_005722 (G3826) の発現量}
 \end{aligned}$$

($D > 0$ のときリンパ節転移あり、 $D \leq 0$ のときリンパ節転移なし、と判別する)。

[0018] 本発明において用いる遺伝子群の解析方法としては、以下のものが挙げられる：

1. Hierarchical Cluster Analysis クラスター分析
2. Logistic Discrimination (変数選択法を含む) ロジスティック解析
3. Classification And Regression Tree カート
4. Principal Component Analysis Artificial Neural Network (拡張法を含む) ピーシーエー／主成分分析
5. Projection Pursuit for supervised classification プロジェクションパーシュート／射影追跡
6. Support Vector Machine サポートベクターマシン／エスブイエム

7. Self Organizing Map エスオーエム

8. AdaBoost アダブースト

9. これらの手法を2つ以上組み合わせた遺伝子の選択の工程。

[0019] 本発明において用いる遺伝子発現情報の解析方法としては、ロジスティック解析 (Logistic Discrimination) が挙げられる。

[0020] 本発明において用いる変数選択の方法としては、以下のものが挙げられる：

1. ステップワイズ (逐次増減法; stepwise)
2. 前進選択法 (forward)
3. 後退選択法 (backward)

発明の効果

[0021] 本発明では、大腸癌患者が大腸癌原発巣切除手術を受ける時点で、大腸癌細胞が周辺のリンパ節に転移している可能性が高いか否かを判定するために有用な一連の遺伝子セット、ならびにそれらの遺伝子発現情報に基づいてリンパ節転移の有無を予測するための判別式が提供される。本発明の方法に従えば、大腸癌原発巣組織の当該遺伝子セットの遺伝子発現情報をロジスティック回帰式で解析することにより、良好なリンパ節転移判定成績を得ることができる。したがって、大腸癌原発巣切除手術の時点において、リンパ節への癌細胞の転移の有無を予測することが可能となる。

発明を実施するための最良の形態

[0022] 本発明の方法は、大腸癌原発巣切除手術の時点におけるリンパ節転移の有無を予測するうえで有効な遺伝子セット、及び当該遺伝子セットの発現量に基づいて実際にリンパ節転移の有無を予測するための判別式によって特徴付けられる。

[0023] リンパ節転移の有無を予測するうえで有用な遺伝子セットは、多サンプルの大腸癌原発巣組織について遺伝子発現を網羅的に調べ、その中から判定に利用可能な遺伝子のセットを選出することにより得られる。このような網羅的な遺伝子発現解析の方法としては、マイクロアレイをはじめとして、Northern解析、ATAC-PCR法 (Katoら、Nuc. Acids Res., vol. 25, p. 4694-4696, 1997) や Taq Man PCR法 (Applied Biosystems社) に代表されるリアルタイムPCR法、SAGE (Velculescuら、Science, vol. 270, p. 48

4-487, 1995)等々様々な方法が利用可能である。

[0024] 本発明の好ましい態様では、DNAマイクロアレイを用いた、特開2004-33082に記載の方法に従って行われる。より具体的には、インフォームドコンセントを経て収集された、原発巣除去手術時の組織病理的観察でリンパ節への転移が認められた患者由来の大腸癌原発巣組織63例と、リンパ節転移が認められなかった患者由来の原発巣組織87例の合計150例の組織について上記DNAマイクロアレイを用いて遺伝子発現データを取得した。比較対照としては、40症例分の大腸癌原発巣組織周辺の正常大腸粘膜組織から得られた遺伝子発現データを用いた。

[0025] 上記の遺伝子発現データは、癌組織から抽出した全RNAを材料として調製した蛍光標識cDNAを、DNAマイクロアレイにハイブリダイズさせ、DNAマイクロアレイ上のプローブにハイブリダイズした蛍光標識cDNAの発する蛍光シグナルを専用のスキャナで検出・定量化することにより取得される。より具体的な手順を以下に記した。

[0026] 大腸癌組織あるいは正常大腸粘膜組織からの全RNA抽出は、TRIzol試薬(GIBCO BRL社)、ISOGEN(ニッポンジーン社)などの試薬を用い、各試薬の添付文書に記載された方法に従って行うことができる。このようにして調製した全RNAは、そのまま下記の標識cDNAの調製に使用することができる。また、例えば、mRNA Purification Kit (Amersham BioSciences社)などの市販のキットにより、添付の方法に従って、該全RNAからポリアデニン付加RNA(以下、「mRNA」と称することもある)を精製して以下の標識cDNAの調製に使用することもできる。

[0027] Cy3標識した大腸癌原発巣組織由来のcDNA(以下、「Cy3cDNA」と称することもある)は、上記の全RNAまたはmRNA、オリゴdTプライマー、dNTP及びCy3標識dUTPを含む混合液に逆転写酵素を加えた後、37~45℃で1~3時間、好ましくは、42℃で1時間加温することにより調製される。比較対照として使用されるCy5標識した正常大腸粘膜由来のcDNA(以下、「Cy5cDNA」と称することもある)の調製も、正常大腸粘膜組織の全RNAを用いて同様の方法により行われる。こうして得られたCy3cDNA及びCy5cDNAは、それぞれ変性溶液中で65~70℃で10~20分間、好ましくは、70℃で10分間加熱処理し、中和後、等量混合される(以下、この混合液を「Cy5・Cy3cDNA」と称することもある)。変性溶液として、50mM EDTAを

含む0.5N NaOH又は1N NaOHなどを用いることができるが、50mM EDTAを含む0.5N NaOHを使用するのが好ましい。Cy5・Cy3cDNAの精製は、例えばMicrocon-30 (Amicon社)などの市販キットを用い、添付の方法に従って行われる。

[0028] Cy5・Cy3cDNAとDNAマイクロアレイにプリントされたプローブとのハイブリダイゼーションは、以下のようにして行われる。まず、プローブを熱変性させるためにDNAマイクロアレイを加熱処理し、これに100℃で2分間加熱処理したCy5・Cy3cDNA含有ハイブリダイゼーション液を滴下し、カバーガラスで覆った後、DNAマイクロアレイを密閉容器に入れ、ハイブリダイゼーションを行う。ハイブリダイゼーション条件としては、ハイブリダイゼーション液がホルムアミドを含む場合には、42℃で12時間以上のハイブリダイゼーションが行われ、ホルムアミドを含まない場合には約68℃で12時間以上のハイブリダイゼーションが行われる。ハイブリダイゼーションの終了後、例えばScan Array 4000 (GSI Lumonics社)などの機器によりCy3とCy5の蛍光をスキャンし、蛍光パターンを画像データとして得る。続いて、これらの画像データを、例えばQuantarrayソフトウェア (GSI Lumonics社)などのマイクロアレイデータ専用解析ソフトを用いて解析することにより、全プローブについてのCy3とCy5の蛍光強度をテキスト形式の数値データとして得ることができる。

[0029] 本発明の好ましい態様では上記DNAマイクロアレイを使用したか、その代わりにハイブリダイゼーションのために有効な鎖長を持つ合成DNAを用いても同様の結果を得ることができる。例えば、本発明で開示された遺伝子名あるいは配列情報に基づいて、その一部の配列からなる約20ヌクレオチド以上の長さを持つ合成DNAをプローブとして、ガラス基盤などに固定化したものを使用することも可能である。

[0030] 一般的に、蛍光強度の低いデータはバックグラウンドの影響を大きく受けているので、例えば蛍光強度が強い方から3,000データポイントだけを残すなどの方法により、蛍光強度の低いプローブのデータは棄却され欠損値として扱われる。続いて、スキャンの際に起こりうるCy3とCy5の検出感度調整のずれを補正して標準化するための操作が行われる。すなわち、各プローブについてのCy3とCy5の蛍光強度値の比である $Cy3/Cy5$ を算出し、底が2の対数値(以下、「 $\log(Cy3/Cy5)$ 」と記載する)に変換し、各プローブについての $\log(Cy3/Cy5)$ 値から、全 $\log(Cy3/Cy5)$ 値の

中央値 (median) を差し引くことにより標準化log (Cy3/Cy5) 値を得ることができる。
該標準化log (Cy3/Cy5) 値は、各遺伝子の発現量として用いることができる。

[0031] このようにして得られた全症例についての標準化された数値データ(以下、「標準化数値データ」と記載することがある)は、一旦統合され、欠損値を多く含むプローブのデータを以降の解析対象から外す目的で次の選択操作が行われる。すなわち、マイクロアレイで解析した全150症例のうちの85%以上にあたる128例以上でデータが取得できているプローブのデータのみが選択される。これにより、欠損値を15%以下しか含まないプローブのデータだけを選択することができる。さらに、個人的な遺伝子背景因子の除外を目的として次の選択操作が加えられる。すなわち、各プローブについて、150例の大腸癌原発巣のデータ内での分散値と12例の正常大腸粘膜についてのデータ内での分散値を算出し、前者が後者の1.1倍を超えているプローブのデータのみが選択される。

[0032] これら一連の選択操作により、2,121種類のプローブの標準化された数値データが以降の解析対象として選択される。このようにして選択される標準化数値データに含まれる欠損値の存在は、後の統計解析において不都合を生じるため、何らかの方法で補完される必要がある。

[0033] 補完の方法としては様々なものが適用可能であるが、例えば、補完する欠損値を含む症例についての全データの平均値に、その欠損値を含む遺伝子の全症例についてのデータの平均値を加えた値から、全症例についての全遺伝子のデータの平均値を引いた値をもって補完する方法がある。他にはTroyanskayaらの報告 (Bioinformatics, vol. 17, p. 520–525, 2001) において3種類の補完方法、すなわち、K-Nearest Neighbors (KNN) method、Singular Value Decomposition (SVD) based method及びrow average methodによる補完の例が示されている。これらのうちのいずれかの方法を適用することにより、全ての欠損値を補完することが可能である。

[0034] かくして準備される欠損値が補完された標準化数値データ(以下、「標準化遺伝子発現データ」と称することもある)は、バックグラウンドの影響を受けておらず、Cy3とCy5の検出感度の違いによる誤差を含まず、また欠損値も含まず、かつ、正常大腸粘膜との比較における大腸癌原発巣の遺伝子発現の変動幅が個人差に起因する遺伝

子発現の変動幅を超えている遺伝子の発現情報を有しており、以後の統計解析の信頼性を確保することができるものである。

[0035] DNAマイクロアレイの測定で得られる大量の遺伝子発現データを統計学的手法により処理し、目的に叶う遺伝子セットを導き出す方法については、確立された一般的なものはなく、研究者の相当な鋭意工夫を必要とするのが現実である。本発明においては、まず4つの異なるアプローチで解析を行い、それらの各々について、リンパ節転移の有無の予測に利用可能な遺伝子群が同定される。

[0036] 本発明で行った4つのアプローチは、

(a) Support Vector Machine (SVM) (Hastieら、The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Springer, 2001)、

(b) Principal Component Analysis/artificial Neural Network (PCA/aNN) (Khanら、Nature Medicine, vol. 7, p. 673 - 679, 2001)の拡張法、

(c) Hierarchical Cluster Analysis (HCA) + Stepwise Logistic Discrimination及び

(d) Classification And Regression Tree (CART) (Breimanら、Classification and Regression Trees, Wadsworth, 1983) + Logistic Discrimination

である。遺伝子群の同定に際しては、統計学的な信頼性を担保する目的で、全データを予測用遺伝子同定用と評価用の2群に分けて解析を行う。より具体的に述べると、150例のデータを、リンパ節転移ありの42例とリンパ節転移なしの57例から成る99例と、リンパ節転移ありの21症例とリンパ節転移なしの30例から成る51例、の2群に分け、前者の99例分のデータをリンパ節転移の有無を予測するための遺伝子の同定と判別式の確立に使用し、その判別式で後者の51例分のデータを判別することにより、判別式の評価を行う。以降の記載においては、予測用遺伝子の同定及び判別式の確立に使用される前者の99例分のデータを「トレーニング用データ」と表現し、判別式の評価に使用する後者の51例分のデータを「テスト用データ」と表現することがある。

[0037] 上記4つのアプローチのうちの(a)、(c)及び(d)の3つに関しては、上記のデータの2分割を、分割時の標本変動を考慮して、ランダムに100回行って解析することにより、100通りの判定用遺伝子群を同定したうえで、同定された回数が多かった遺伝

子を採用する。

- [0038] 一方、上記(b)のアプローチに関しては、膨大な計算量を考慮して最初の2分割についてのみ実施する。ただし、トレーニング用の99例分のデータを2対1の割合で1250回ランダムに2分割し、それを用いて主成分分析とニューラルネットワークの学習を反復する。学習後、リンパ節転移の有無を識別する感度に基づき遺伝子をランキングし、遺伝子を絞り込む。2121個の遺伝子から開始し、以降1536個、768個、384個、192個、96個、48個、24個の絞り込み個数のそれぞれで学習を進める。
- [0039] 以上の解析を行うことにより、各遺伝子セットに含まれる遺伝子の個数と、確立された判別式を用いたテスト用データの正分類率(判別式での判別結果と組織病理学的検査の結果が一致した症例数/テスト用データ数×100(%))の平均値として、(a)については遺伝子数が144個で正分類率は80.2%(標準偏差は5.6%)、(b)については遺伝子数が192個で正分類率は(90.2%)、(c)については遺伝子数が133個で正分類率は78.6%(標準偏差は6.2%)及び(d)については遺伝子数が138個で正分類率は86.3%(標準偏差:4.5%)が得られる。このとき、16種類の遺伝子が、各アプローチで選択された遺伝子セットに共通して含まれる。
- [0040] 次いで、正分類率を落とさないようにしつつ、予測に使用する遺伝子の数を絞り込むために、まず対象とする遺伝子を上記の16遺伝子とし、これら16遺伝子各々の寄与(以下、「主効果」と記載する)に加えて、2遺伝子の交互作用も加味した統計解析を行う。これにより、個別の遺伝子による主効果だけでなく、遺伝子間の交互作用を含めたより広い範囲で判別ルールを探索することとなり、高い判別性能を維持できることが期待される。
- [0041] 交互作用探索のために、上記解析で用いた100通りのトレーニング用データのそれぞれで、リンパ節転移の有無を応答とするCART解析を再度行う。このCART解析では、FreeのソフトウェアRのrpartを用いた。その際、操作パラメータは全てデフォルトの値を用いた。この解析により、データの分割を指示する変数として登場する遺伝子の個数として、1回の解析あたり3個から5個が得られる。
- [0042] 登場する遺伝子が例えば3個の場合、とり得る遺伝子のペアは3通りあるため、それら全てのペアを交互作用として捉える。同様にして、遺伝子が4個の場合は6組、5個

の場合は10組の各ペアを交互作用として捉える。そして、できる限り多くの候補をカバーするため100通りの解析のうち12回以上現れた18組の遺伝子ペアを交互作用の候補として選択する。

- [0043] 次に判別式の確立のために、150例のデータを用いて、16遺伝子の主効果と18通りの交互作用を説明変数とし、リンパ節転移の有無を応答としたロジスティック回帰モデルにおいてステップワイズの変数選択を行う。その際、回帰係数の有意性検定のp値を、変数の組入れ基準(0.05未満)及び除外基準(0.05超)として用いる。これにより、6個の変数、すなわち、G1592、G3031、G3826、G4370、G2645とG3177の交互作用、G3753とG3826の交互作用が選択される。そして、リンパ節転移の有無を予測するための判別式は、

$$\begin{aligned} D = & 0.2307 - 2.7132 \times \text{「G1592の発現量」} \\ & + 8.9509 \times \text{「G3031の発現量」} \\ & + 8.7975 \times \text{「G3826の発現量」} \\ & - 2.3098 \times \text{「G4370の発現量」} \\ & + 3.5126 \times \text{「G2645の発現量」} \times \text{「G3177の発現量」} \\ & - 8.8226 \times \text{「G3753の発現量」} \times \text{「G3826の発現量」} \end{aligned}$$

と推定され、 $D > 0$ のときリンパ節転移あり、 $D \leq 0$ のときリンパ節転移なし、とする判別ルールが導かれる。この判別式に登場した7個の遺伝子、すなわちG1592、G2645、G3031、G3177、G3753、G3826、G4370をリンパ節転移の有無の識別に寄与する遺伝子のセットとして選択する。それらの遺伝子名を表1に記した。

- [0044] 最後に、選択した遺伝子セットによるリンパ節転移の判別性能がLOO法により評価される。すなわち、1サンプルを除いた残りの149サンプルのデータを用いて、上記の6個の変数を含むロジスティック判別式を推定し、それによって除いたサンプルを判別する操作を、150サンプルのそれぞれで実施する。これにより、表2に示すように、選択した遺伝子のセットによる正分類率は88.7%(感度:77.8%、特異度:96.6%)と推定される。以上のように、本発明においては、大腸癌のリンパ節転移の有無を高い精度で予測するのに必要な遺伝子セットを明らかにすることができる。

- [0045] 以下に本発明を実施例により詳細に説明するが、これら実施例によって本発明は

何ら制約を受けることはない。なお、実施例において使用した試薬類は特にことわりのない限り、ナカライテスク株式会社より購入したものを使用した。

実施例 1

[0046] (1) 大腸癌組織試料からの全RNA調製

DNAマイクロアレイを用いた、大腸癌における遺伝子発現解析を行うための試料としては、インフォームドコンセントを経て収集された、大腸癌手術時に切除された大腸癌原発巣組織150例を用いた。その内訳は、原発巣除去手術時の組織病理学的な観察でリンパ節転移が認められた患者に由来する63例(以下、「リンパ節転移陽性症例」と記載する)と、リンパ節転移が認められなかった患者に由来する87例(以下、「リンパ節転移陰性症例」と記載する)である。これらの大腸癌組織試料からTRIzol試薬(GIBCO BRL社より購入)を用いて全RNAを抽出した。抽出手順は基本的に上記試薬に添付のマニュアルに従った。この他に、40例分の正常大腸粘膜部分由来の全RNAを抽出し、それらを混合して、全ての実験を通して使用する標準正常大腸粘膜全RNAとした。これらのRNAサンプルの濃度は、定法通りに分光光度計を用いて測定した波長260nmでの吸光度に基づいて算出した。

[0047] (2) 蛍光ラベルターゲットの調製

DNAマイクロアレイにハイブリダイズさせる蛍光ラベルターゲットは以下の手順で作製した。まず、25 μ gの大腸癌部試料由来全RNA(以下、「大腸癌RNA」と記す)と25 μ gの標準正常大腸粘膜全RNA(以下、「標準大腸粘膜RNA」と記す)を別々のチューブに入れ、それぞれに2 μ gの18ヌクレオチドから成るオリゴdTプライマーを加え、滅菌蒸留水にて容量を14 μ Lとし、70°Cで10分間加熱した後、直ちに氷上に移して急冷した。その後、それぞれのチューブに、6 μ Lの5×First Strand Buffer、3 μ Lの0.1M DTT、1.5 μ Lの20×dNTPmix(10mMのdATP、dCTP、dGTP及び6mMのdTTPの混合物)及び0.5 μ LのRNAGuardを添加した。

[0048] さらに、大腸癌RNAを入れた方のチューブに蛍光色素Cy3でラベルされたdUTP(以下、「Cy3-dUTP」と記す;濃度1mM)を3 μ L、標準大腸粘膜RNAを入れた方のチューブにCy5でラベルされたdUTP(以下、「Cy5-dUTP」と記す;濃度1mM)を3 μ L加えて、42°Cにて2分間保温した。その後、逆転写酵素であるSuperscrip

tIIを各チューブに2 μ L加えて、42°Cにてさらに1時間保温することによりラベル反応を行った。この反応により、大腸癌RNAと標準大腸粘膜RNAを鋳型としてcDNA合成が起こる際に、それぞれCy3-dUTPとCy5-dUTPが取り込まれることにより、それぞれCy3とCy5で蛍光ラベルされた大腸癌ラベルターゲットと標準大腸粘膜ラベルターゲットが生成する。

[0049] この反応で使用した5×First Strand Buffer、0.1M DTT及びSuperscriptIIは、いずれもGIBCO BRL社より購入した。また、dATP、dCTP、dGTP及びdTTP、Cy5-dUTP及びCy3-dUTP、そしてRNAguardはいずれもAmersham Biosciences社より購入した。反応後は、各チューブに5 μ Lの変性溶液(0.5N NaOH、50mM EDTA)を添加して70°Cで10分間加熱した後、7.5 μ Lの1M Tris-HCl(pH7.5)を加えることにより中和した。これらの処理を行った段階で、大腸癌ラベルターゲットと標準大腸粘膜ラベルターゲットを混合し、ここに10 μ gのヒトCOT-1 DNA(GIBCO BRL社より購入)を添加した。この混合液にTEバッファーを加えて500 μ Lに調整し、Microcon-30(Amicon社より購入)を用いて精製・濃縮することにより、未反応のCy5-dUTP及びCy3-dUTPなどを除去した。精製・濃縮の手順はMicrocon-30に添付のマニュアルに従った。最終的には、全容量が5 μ Lとなるまで濃縮し、これをDNAマイクロアレイにハイブリダイズさせるラベルターゲットとした。

[0050] (3)DNAマイクロアレイの前処理

DNAマイクロアレイをマスキング溶液(3gの無水コハク酸、190mLのN-メチル-2-ピロリドン及び21mLの0.2Mホウ酸ナトリウムの混合液)に5分間浸すことによりマスキングを行った後、95°Cの蒸留水に3分間浸すことにより、マイクロアレイ上にプリントされているcDNAを熱変性させた。その後直ちに95%以上のエタノールに1分間浸して脱水し風乾させた。

[0051] (4)ラベルターゲットとDNAマイクロアレイとのハイブリダイゼーション

前述のようにして調製したラベルターゲット溶液5 μ Lに対して、2.5 μ Lの10mg/mLのポリアデニン(Roche社より購入)、0.5 μ Lの10%SDS溶液、3 μ Lの20×PM溶液(0.4%BSAと1%SDSの混合液)、15 μ Lのホルムアミド、3 μ Lの20×SSC(3M塩化ナトリウム、0.3Mクエン酸ナトリウム、pH7.0)及び滅菌蒸留水1 μ Lを添加

し、100℃で2分間加熱した後、暗所にて約30分間室温で静置した。その後、前項に記載の方法で前処理したDNAマイクロアレイのcDNAがプリントされている部分に滴下し、24×40ミリメートルのカバーガラス(マツナミガラス工業より購入)で覆い、マイクロアレイを密閉容器に入れ、その容器ごと42℃のインキュベーターに約16時間入れておくことにより、ラベルターゲットをマイクロアレイ上のcDNAにハイブリダイズさせた。ハイブリダイゼーションの後、マイクロアレイを0.1%SDSを含む2×SSCに浸して10分間洗浄し、次に、0.1%SDSを含む0.1×SSCに浸して10分間洗浄した。さらに、0.1×SSCに浸して5分間の洗浄を2回行った後、滴を切って暗所で風乾させた。

[0052] (5) マイクロアレイのスキャンとデータ解析

洗浄後風乾させたマイクロアレイを、マイクロアレイ専用共焦点レーザースキャナであるScanArray 4000 (GSI Lumonics社製)を使ってCy3とCy5の蛍光を独立にスキャンすることにより、マイクロアレイ上の各プローブにハイブリダイズした大腸癌ターゲットと標準大腸ターゲットに由来するCy3とCy5の蛍光パターンを16ビットTiff形式のスキャン画像データとして得た。続いて、それらの画像データをマイクロアレイデータ専用解析ソフトであるQuantArrayソフトウェア (GSI Lumonics社製)を用いて解析することにより、全プローブについてのCy3とCy5の蛍光強度をテキスト形式の数値データとして得た。バックグラウンドの補正のために、cDNAがプリントされていない部分の蛍光強度値を、各プローブについての蛍光強度値から差し引いた。また、蛍光強度値が低い部分は実験誤差の影響を大きく受けるため、蛍光強度値が高い方から約3000のデータポイントを残して他のデータは棄却した。各プローブについてのCy3とCy5の蛍光強度値の比、すなわち $Cy3/Cy5$ を算出し、底が2の対数値(以下、「 $\log(Cy3/Cy5)$ 」と記載する)に変換した。スキャンの際に起こりうるCy3とCy5の検出感度調整のずれを補正して標準化するために、各プローブについての $\log(Cy3/Cy5)$ 値から、全 $\log(Cy3/Cy5)$ 値の中央値(median)を差し引くことにより標準化 $\log(Cy3/Cy5)$ 値を得た。

[0053] 以上の操作により、標準大腸粘膜RNAを基準としたときの、リンパ節転移ありの症例63例分及びリンパ節転移なしの症例87例分の大腸癌原発巣の相対的発現強度

を対数化し、標準化した数値データを得ることができた。また、同様の操作によって、標準大腸粘膜RNAを基準としたときの、正常大腸粘膜サンプル12例分の数値データも得た。これらの数値データのうち、解析した150症例の大腸癌原発巣のうちの85%にあたる128症例以上についてデータが取得できており、かつ、150症例の大腸癌原発巣のデータ内での分散値 (variance) が、12例の正常大腸粘膜についてのデータ内での分散値の1.1倍を超えていた合計2,121種類のプローブについてのデータのみを以降の統計解析に使用した。

- [0054] これらの2,121プローブのデータの中には、症例によっては発現量が低く、カットオフ値を下回ったために棄却されたものも含まれており、それらのデータは欠損値となっている。このような欠損値は合計で8,816個存在していた。全データ数は $150 \times 2,121 = 318,150$ 個であることから、欠損値の含有率は約2.8%である。これらの欠損値の存在は、後の統計解析において不都合を生じるため、k-Nearest Neighbor法を用いて、それらを補完した。具体的には、150症例 $\times$ 2,121プローブのデータ行列において、欠損値となっていた発現量を、サンプル間距離でその遺伝子に最も近い8個の遺伝子の発現量の平均値で推定した。具体的には、150症例 $\times$ 2,121プローブのデータ行列において、ペアにした全てのサンプル間の距離を計算した。そして、欠損値のあるサンプルと最も距離の近かった8個のサンプルで補完すべき遺伝子発現量の平均値を求め、それを欠損値の補完値として用いた。ここに、欠損値のあるサンプルと最も距離の近いサンプルの個数は、個数を順次増加させroot mean square (RMS) が最小となる個数で定義した。このようにして欠損値を補完して得られた全数値データを以降、標準化遺伝子発現データと記載する。

実施例 2

- [0055] 標準化遺伝子発現データの統計解析によるリンパ節転移予測のための高判別能遺伝子セットの決定

本実施例においては、DNAマイクロアレイにプリントしたプローブを指して遺伝子と呼称することがある。

最初の試みとして、判定用遺伝子セットの同定のために次の4つのアプローチを適用した。すなわち、(a) Support Vector Machine (SVM) (Hastieら、The Elements of St

atistical Learning–Data Mining, Inference, and Prediction, Springer, 2001)、(b) Principal Component Analysis/artificial Neural Network (PCA/aNN) (Khanら、Nature Medicine, vol. 7, p. 673–679, 2001)の拡張法、(c) Hierarchical Cluster Analysis (HCA) + Stepwise Logistic Discrimination及び(d) Classification And Regression Tree (CART) (Breimanら、Classification and Regression Trees, Wadsworth, 1983) + Logistic Discrimination、の4通りの手法を用いた。

[0056] 遺伝子群の同定に際しては、統計学的な信頼性を担保する目的で、全データを判定用遺伝子同定用と評価用の2群に分けて解析を行った。より具体的に述べると、150例のデータを、リンパ節転移ありの42例とリンパ節転移なしの57例から成る99例と、リンパ節転移ありの21症例とリンパ節転移なしの30例から成る51例、の2群に分け、前者の99例分のデータをリンパ節転移の有無を判定するための遺伝子の同定と判別式の確立に使用し、その判別式で後者の51例分のデータを判別することにより、遺伝子の同定と判別式の評価を行った。以降の記載においては、前者の99例分のデータのように判定用遺伝子の同定及び判別式の確立に使用されるデータを「トレーニング用データ」と表現し、後者の51例分のデータのように判別式の評価に使用するデータを「テスト用データ」と表現することがある。

[0057] 上記4つのアプローチのうちの(a)、(c)及び(d)の3つに関しては、上記のデータの2分割をランダムに100回行って解析することにより、100通りの判定用遺伝子群を同定したうえで、同定された回数が多かった遺伝子を採用した。一方、上記(b)のアプローチに関しては、最初の2分割についてのみ実施した。ただし、トレーニング用の99例分のデータを2対1の割合で1250回ランダムに2分割し、それを用いて主成分分析とニューラルネットワークの学習を反復した。学習後、リンパ節転移の有無を識別する感度に基づき遺伝子をランキングし、遺伝子を絞り込んだ。2121個の遺伝子から開始し、以降1536個、768個、384個、192個、96個、48個、24個の絞り込み個数のそれぞれで学習を進めた。

[0058] 以上の検討の結果、各アプローチごとに判別用遺伝子セットと判別式を確立することができた。各遺伝子セットに含まれる遺伝子の個数と、確立された判別式を用いたテスト用データの正分類率(判別式での判別結果と組織病理学的検査の結果が一

致した症例数／テスト用データ数×100(%))の平均値は、(a)については遺伝子数が144個で正分類率は80.2%(標準偏差は5.6%)、(b)については遺伝子数が192個で正分類率は(90.2%)、(c)については遺伝子数が133個で正分類率は78.6%(標準偏差は6.2%)及び(d)については遺伝子数が138個で正分類率は86.3%(標準偏差:4.5%)であった。また、各アプローチで選択された遺伝子セットに共通して含まれていた遺伝子が16種類あった。

[0059] 上記のように、4つの異なるアプローチのそれぞれで、80%を超える高い正分類率を達成することができた。しかしながら、判別に使用する遺伝子の数はどのアプローチでも100種類を超えており、実際にアッセイ法として実用化するためには多すぎると考えた。

[0060] そこで、正分類率を落とさないようにしつつ、判定に使用する遺伝子の数を絞り込んで新たな判別ルールを確立することを考えた。そのために、まず対象とする遺伝子を上記の各アプローチで選択された遺伝子セットに共通して含まれていた16遺伝子とした。次に、これら16遺伝子各々の寄与(以下、「主効果」と記載する)に加えて、2遺伝子の交互作用も加味することに工夫した。個別の遺伝子による主効果だけでなく、遺伝子間の交互作用を含めたより広い範囲で判別ルールを探索することで、高い判別性能を維持できることを期待した。

[0061] 交互作用探索のために、前述の解析で用いた100通りのトレーニング用データのそれぞれで、リンパ節転移の有無を応答とするCART解析を再度行った。その結果、データの分割を指示する変数として登場した遺伝子の個数は、1回の解析あたり3個から5個であった。登場した遺伝子が例えば3個の場合、とり得る遺伝子のペアは3通りあるため、それら全てのペアを交互作用として捉えた。同様にして、遺伝子が4個の場合は6組、5個の場合は10組の各ペアを交互作用として捉えた。そして、100通りの解析のうち、12回以上現れた18組の遺伝子ペアを交互作用の候補として選択した。

[0062] 次に判別式の確立のために、150例のデータを用いて、16遺伝子の主効果と18通りの交互作用を説明変数とし、リンパ節転移の有無を応答としたロジスティック回帰モデルにおいてステップワイズの変数選択を行った。その際、回帰係数の有意性検

定のp値を、変数の組入れ基準(0.05未満)及び除外基準(0.05超)として用いた。その結果、6個の変数、すなわち、G1592、G3031、G3826、G4370、G2645とG3177の交互作用、G3753とG3826の交互作用が選択された。G1592、G3031、G3826、G4370、G2645、G3177及びG3753とは、本発明で使用したColonoChip上の各プローブ(遺伝子)に付与したシリアル番号である。そして、リンパ節転移の有無を判定するための判別式は、

$$\begin{aligned} D = & 0.2307 - 2.7132 \times \text{G1592の発現量} \\ & + 8.9509 \times \text{G3031の発現量} \\ & + 8.7975 \times \text{G3826の発現量} \\ & - 2.3098 \times \text{G4370の発現量} \\ & + 3.5126 \times \text{G2645の発現量} \times \text{G3177の発現量} \\ & - 8.8226 \times \text{G3753の発現量} \times \text{G3826の発現量} \end{aligned}$$

と推定され、 $D > 0$ のときリンパ節転移あり、 $D \leq 0$ のときリンパ節転移なし、とする判別ルールが導かれた。

[0063] この判別式に登場した7個の遺伝子、すなわちG1592、G2645、G3031、G3177、G3753、G3826、及びG4370をリンパ節転移の有無の識別に寄与する遺伝子のセットとして選択した。それらの遺伝子名を表1に記した。表1中には上記の各遺伝子のRefSeqデータベースにおけるアクセス番号も併記した。RefSeqデータベースにはNational Center for Biotechnology Information (NCBI) のWeb site (<http://www.ncbi.nlm.nih.gov/RefSeq/index.html>) からアクセスすることができる。

[0064] [表1]

シリアル番号	遺伝子名	RefSeqアクセス番号
G1592	チロシン 3-モノオキシゲナーゼ/トリプトファン 5-モノオキシゲナーゼ活性化蛋白、 β ポリペプチド	NM_003404
G2645	high-mobility group box 1	NM_002128
G3031	免疫グロブリンスーパーファミリー、member 8	NM_052868
G3177	RNAポリメラーゼII、ポリペプチドK	NM_005034
G3753	熱ショック27kD蛋白1	NM_001540
G3826	ARP2アクチン関連蛋白2	NM_005722
G4370	マウス 1 a 関連蛋白のヒト様ortholog	NM_015315

[0065] 最後に、選択した遺伝子セットによるリンパ節転移の判定性能をLOO法により評価

した。すなわち、1サンプルを除いた残りの149サンプルのデータを用いて、上記の6個の変数を含むロジスティック判別式を推定し、それによって除いたサンプルを判別する操作を、150サンプルのそれぞれで実施した。その結果を表2に示した。表2から、選択した遺伝子のセットによる正分類率は88.7% (感度:77.8%、特異度:96.6%)と推定された。

[0066] [表2]

予測 実際	転移あり	転移なし	計
転移あり	4 9	1 4	6 3
転移なし	3	8 4	8 7
計	5 2	9 8	1 5 0

産業上の利用可能性

[0067] 本発明により可能となるリンパ節転移の判定によって、症例に応じたよりよい治療方針の選択が可能であり、また医療経済効果も期待できる。例えば、リンパ節転移の可能性の高い症例に対しては積極的な治療を施行することによって予後の改善が期待できるし、一方、リンパ節転移の可能性の低い症例に対しては、術後補助療法は軽度のものにすることができ、患者の肉体的・経済的負担を軽くすることができる。

[0068] さらに、本発明で開示される遺伝子セットに含まれる個々の遺伝子は、リンパ節転移の原因として機能するものである可能性も推察されるため、これらの遺伝子及びその発現産物を標的とする薬剤を開発し、リンパ節転移を直接抑制できるようにすることも期待できる。

請求の範囲

- [1] 下記(1)～(4)の工程を含む、大腸癌リンパ節転移の有無を予測するための遺伝子セットの選択方法：
- (1)組織病理学的判定によりリンパ節転移の有無が明らかにされた患者の大腸癌原発巣組織における遺伝子発現情報を、教師あり学習解析方法を少なくとも一つ含む、4以上の解析方法で解析することにより、リンパ節転移の有無を正分類率75%以上で分類できる遺伝子群をそれぞれの解析方法において選定する工程、
- (2)(1)で用いたそれぞれの解析方法で選定された遺伝子群から、何れの解析方法でも共通して選定された共通遺伝子を選択する工程、
- (3)前記遺伝子発現情報を解析することにより、任意の2以上の遺伝子の組合せの中から、リンパ節転移の有無の分類を指示し、交互作用を示す遺伝子の組合わせを選択する工程、及び
- (4)前記共通遺伝子及び前記遺伝子の組合わせを説明変数として、リンパ節転移の有無を応答としたロジスティック回帰モデルにおける変数選択を行う工程。
- [2] (1)の解析方法が、(a)Support Vector Machine、(b)Principal Component Analysis Artificial Neural Networkの拡張法、(c)Hierarchical Cluster AnalysisとStepwise Logistic Discriminationの組合せ及び(d)Classification And Regression TreeとLogistic Discriminationの組合せよりなる群から選択されるものを少なくとも一つ含むものである、請求項1に記載の方法。
- [3] (3)の解析方法が、(d)Classification And Regression TreeとLogistic Discriminationの組合せである、請求項1または2に記載の方法。
- [4] (4)の変数選択の方法が、ステップワイズの変数選択法である、請求項1ないし3のいずれかに記載の方法。
- [5] 請求項1ないし4のいずれかに記載の方法により選択される、大腸癌リンパ節転移の有無を予測するための遺伝子セット。
- [6] 少なくともNM_003404(G1592)、NM_002128(G2645)、NM_052868(G3031)、NM_005034(G3177)、NM_001540(G3753)、NM_005722(G3826)、及びNM_015315(G4370)のデータベースのアクセス番号(シリアル番号)で表される遺伝子を含む、請求項5

に記載の遺伝子セット。

[7] 請求項5または6の何れかに記載の遺伝子セットを用いることを特徴とする大腸癌リンパ節転移の有無を予測するための方法。

[8] 下記の判別式を用いることを特徴とする請求項7記載の方法：

$$\begin{aligned} D = & 0.2307 - 2.7132 \times \text{NM\_003404 (G1592) の発現量} \\ & + 8.9509 \times \text{NM\_052868 (G3031) の発現量} \\ & + 8.7975 \times \text{NM\_005722 (G3826) の発現量} \\ & - 2.3098 \times \text{NM\_015315 (G4370) の発現量} \\ & + 3.5126 \times \text{NM\_002128 (G2645) の発現量} \times \text{NM\_005034 (G3177) の発現量} \\ & - 8.8226 \times \text{NM\_001540 (G3753) の発現量} \times \text{NM\_005722 (G3826) の発現量} \end{aligned}$$

($D > 0$ のときリンパ節転移あり、 $D \leq 0$ のときリンパ節転移なし、と判別する)。

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315143

A. CLASSIFICATION OF SUBJECT MATTER

C12Q1/68(2006.01) i, C12N15/09(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q1/68, C12N15/09

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

BIOSIS/MEDLINE/WPIDS (STN), CAPLUS (STN), JSTPLUS/JMEDPLUS (JDream2)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KOMORI, T. et al., "Gene expressin profiling for the prediction of lymph node metastasis in colorectal cancer", The Molecular Biology Society of Japan(2004), Vol.27, pp.1019, full text	1-8
Y	BOULESTEIX, A.-L. et al., A CART-based approach to discover emerging patterns in microarray data", BIOINFORMATICS, (2003), Vol.19, No.18, pp.2465-2472, full text	1-8

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
20 October, 2006 (20.10.06)

Date of mailing of the international search report
07 November, 2006 (07.11.06)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315143

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YANG, X. et al., "The relationship between expression of c-ras, c-erbB-2, nm23, and p53 gene products and development of trophoblastic tumor and their predictive significance for the malignant transformation of complete hydatidiform mole", Gynecologic Oncology, (2002), Vol.85, pp.438-444, full text	1-8
A	ARTINYAN, A. et al., "Increased risk of lymph node metastasis in T3 colon cancers with decreased thymidylate synthase expression" Gastroenterology, (APR. 2005), Vol.128, No.4, Suppl.2, pp. A784, full text	1-8
A	JP 2004-33082 A (Ichiro TAKEMASA), 05 February, 2004 (05.02.04), Full text (Family: none)	1-8
A	COFFEY, C. S. et al., "An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation.", BMC Bioinformatics, (2004), Vol.5, 49, full text	1-8
A	MASAKI, T. et al., "Expression of adhesion molecules as a novel risk factor for lymph node metastasis or local recurrence in early invasive (T1) colorectal carcinomas", Journal of the Japan Society of Coloproctology, (2002), Vol.55, No.10, pp.858-866, full text	1-8
A	ORIAN-ROUSSEAU, V. et al., "Genes upregulated in a metastasizing human colon carcinoma cell line", Int. J. Cancer, (Feb. 2005), Vol.113, pp.699-705, full text	1-8
A	KUNIYASU, H. et al., Co-expression of receptor for advanced ligation end products and the ligand amphoterin associates closely with metastasis of colorectal cancer" Oncology reports, (2003) Vol. 10, No.2, pp.445-448, full text	1-8
A	OTSUBO, T. et al., "Involvement of Arp2/3 complex in the process of colorectal carcinogenesis", Modern Pathology, (2004), Vol.17, pp.461-467, full text	1-8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315143

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2005/054508 A2 (IPSOGEN), 16 June, 2005 (16.06.05), Full text (particularly, Gene No.431, 547, pp111-113) & US 2005/287544 A1	1-8
A	KOMURO, K. et al., "Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis J.Surgical Res., (APR. 2005), Vol.124, pp.216-224, full text	1-8
A	WO 2005/028676 A2 (ONCOTHERAPY SCIENCE, INC.), 31 March, 2005 (31.03.05), Full text (particularly, Table.3, BRC No.171) & EP 1668156 A2	1-8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315143

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

The inventions of claims 1-8 are classified into a general inventive concept relating to "a method for selecting a set of genes for use in the prediction of the presence or absence of lymph node metastasis of colorectal cancer" (claims 1-5 and a part of claim 7) and a general inventive concept relating to "a set of genes including a gene indicated by an accession number in a specific database" (claims 6 and 8 and a part of claim 7).

Thus, claims 1-8 include two general inventive concepts relating to "a set of genes for use in the prediction of the presence or absence of lymph node metastasis of colorectal cancer". However, there is no common novel
(continued to extra sheet)

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest
the

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, payment of a protest fee..
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☒ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2006/315143

Continuation of Box No.III of continuation of first sheet (2)

special technical feature between the two general inventive concepts. Consequently, the present international application cannot be regarded as satisfying the requirement of unity of invention (PCT Rule 13 (13.1, 13.2 and 13.3)).

A. 発明の属する分野の分類（国際特許分類（I P C）） Int.Cl. C12Q1/68(2006.01)i, C12N15/09(2006.01)i			
B. 調査を行った分野 調査を行った最小限資料（国際特許分類（I P C）） Int.Cl. C12Q1/68, C12N15/09			
最小限資料以外の資料で調査を行った分野に含まれるもの			
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語） BIOSIS/MEDLINE/WPIDS(STN) CAplus(STN) JSTPlus/JMEDPlus(JDream2)			
C. 関連すると認められる文献			
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号	
Y	KOMORI, T. et al., "Gene expressin profiling for the prediction of lymph node metastasis in colorectal cancer", 日本分子生物学会年会プログラム, (2004), Vol.27, pp.1019, 全文	1-8	
Y	BOULESTEIX, A.-L. et al., "A CART-based approach to discover emerging patterns in microarray data", BIOINFORMATICS, (2003), Vol.19, No.18, pp.2465-2472, 全文	1-8	
☒ C 欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。			
* 引用文献のカテゴリー 「A」特に関連のある文献ではなく、一般的技術水準を示すもの 「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの 「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す） 「O」口頭による開示、使用、展示等に言及する文献 「P」国際出願日前で、かつ優先権の主張の基礎となる出願		の日の後に公表された文献 「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの 「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの 「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの 「&」同一パテントファミリー文献	
国際調査を完了した日 2 0 . 1 0 . 2 0 0 6		国際調査報告の発送日 0 7 . 1 1 . 2 0 0 6	
国際調査機関の名称及びあて先 日本国特許庁（I S A / J P） 郵便番号100-8915 東京都千代田区霞が関三丁目4番3号		特許庁審査官（権限のある職員） 佐藤 巖	4 B 3 3 3 4
		電話番号 03-3581-1101 内線 3448	

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求の範囲の番号
Y	YANG, X. et al., "The relationship between expression of c-ras, c-erbB-2, nm23, and p53 gene products and development of trophoblastic tumor and their predictive significance for the malignant transformation of complete hydatidiform mole", Gynecologic Oncology, (2002), Vol.85, pp.438-444, 全文	1-8
A	ARTINYAN, A. et al., "Increased risk of lymph node metastasis in T3 colon cancers with decreased thymidylate synthase expression" Gastroenterology, (APR. 2005), Vol.128, No.4, Suppl.2, pp. A784, 全文	1-8
A	JP 2004-33082 A, (竹政 伊知郎), 2004.02.05, 全文 (ファミリーなし)	1-8
A	COFFEY, C. S. et al., "An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation.", BMC Bioinformatics, (2004), Vol.5, 49, 全文	1-8
A	MASAKI, T. et al., "Expression of adhesion molecules as a novel risk factor for lymph node metastasis or local recurrence in early invasive (T1) colorectal carcinomas", Journal of the Japan Society of Coloproctology, (2002), Vol.55, No.10, pp.858-866, 全文	1-8
A	ORIAN-ROUSSEAU, V. et al., "Genes upregulated in a metastasizing human colon carcinoma cell line", Int. J. Cancer, (Feb. 2005), Vol.113, pp.699-705, 全文	1-8
A	KUNIYASU, H. et al., "Co-expression of receptor for advanced glycation end products and the ligand amphoterin associates closely with metastasis of colorectal cancer" Oncology reports, (2003) Vol. 10, No.2, pp.445-448, 全文	1-8
A	OTSUBO, T. et al., "Involvement of Arp2/3 complex in the process of colorectal carcinogenesis", Modern Pathology, (2004), Vol.17, pp.461-467, 全文	1-8
A	WO 2005/054508 A2, (IPSOGEN), 2005.06.16, 全文(特に、Gene No.431, 547, pp111-113 参照。) & US 2005/287544 A1	1-8
A	KOMURO, K. et al., "Right- and left-sided colorectal cancers display distinct expression profiles and the anatomical stratification allows a high accuracy prediction of lymph node metastasis", J. Surgical Res., (APR. 2005), Vol.124, pp.216-224, 全文	1-8
A	WO 2005/028676 A2, (ONCOTHERAPY SCIENCE, INC.), 2005.03.31, 全文(特に、Table.3, BRC No.171 参照。) & EP 1668156 A2	1-8

第Ⅱ欄 請求の範囲の一部の調査ができないときの意見（第1ページの2の続き）

法第8条第3項（P C T 17条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1. ☐ 請求の範囲 _____ は、この国際調査機関が調査をすることを要しない対象に係るものである。つまり、
2. ☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、
3. ☐ 請求の範囲 _____ は、従属請求の範囲であってP C T規則6.4(a)の第2文及び第3文の規定に従って記載されていない。

第Ⅲ欄 発明の単一性が欠如しているときの意見（第1ページの3の続き）

次に述べるようにこの国際出願に二以上の発明があるところの国際調査機関は認めた。

請求の範囲1-8に記載の発明は、「大腸癌リンパ節転移の有無を予測するための遺伝子セットの選択方法」の一般的発明概念(請求の範囲1-5と7の一部)と、「特定のデータベースのアクセス番号で表された遺伝子を含む遺伝子セット」の一般的発明概念(請求の範囲6,8と7の一部)の、2つに分類される。

よって、請求の範囲1-8には「大腸癌リンパ節転移の有無を予測するための遺伝子セット」に係る2の一般的発明概念が記載されているが、それぞれの一般的発明概念の間には、新規な特別の技術的特徴が共有されていないため、この国際出願は発明の単一性の要件（法施行規則第13条（P C T規則13.1、13.2及び13.3））を満たしていないと認める。

1. ☒ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。
2. ☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。
3. ☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったため、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。
4. ☐ 出願人が必要な追加調査手数料を期間内に納付しなかったため、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意

- ☐ 追加調査手数料及び、該当する場合には、異議申立手数料の納付と共に、出願人から異議申立てがあった。
- ☐ 追加調査手数料の納付と共に出願人から異議申立てがあったが、異議申立手数料が納付命令書に示した期間内に支払われなかった。
- ☒ 追加調査手数料の納付を伴う異議申立てがなかった。