様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ３ 日現在

機関番号：１３８０２
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１５６２９

研究課題名（和文）データマイニングを活用した遺伝子型-表現型解析手法の確立

研究課題名（英文）A datamining approach for genotype-phenotype correlation analysis

研究代表者
吉田　秀一（Yoshida, Shuichi）

浜松医科大学・医学部・特任助教

研究者番号：１０５８０５７４

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：本研究は、これまでに公共の生命科学系データベースなどに蓄積されている様々なゲノムデータを活用したデータマイニングによる遺伝子型-表現型相関モデルの構築と病的意義の大きさを定量的に見積もる統計学的手法の構築を目的とした。モデル疾患を対象にデータマイニングによる遺伝子型-表現型相関の解析を行い、得られたパターンを予測因子として表現型予測モデルを構築し、クロスバリデーション法により予測精度指標を算出・評価したところ、本提案手法において高い予測精度が得られた。今後は、本提案手法の他疾患への応用を検討する。

研究成果の学術的意義や社会的意義
本研究における統計学的手法による遺伝型－表現型相関解析および表現型予測モデルの構築は、将来的に遺伝性疾患の原因となる病的バリアントの探索におけるスクリーニングや遺伝子診断における疾患感受性遺伝子予測等への応用が期待される。

研究成果の概要（英文）：In this study, we proposed a datamining approach using public life science databases for genotype-phenotype correlation analysis. We performed to test our phenotype prediction model for dataset of model disease by cross-validation method. In the result, our prediction model obtained high prediction accuracy.

研究分野：バイオインフォマティクス

キーワード：遺伝型-表現型予測　データマイニング　遺伝子診断

１．研究開始当初の背景

遺伝性疾患の約 85%は、遺伝子中のタンパク質コーディング領域のバリアントが原因であると推察されている。加えて、次世代シークエンサ（NGS）の登場により、遺伝性疾患の責任遺伝子解析が急速に進展している。その一方で、数ある候補バリアントの中から病的意義の強いバリアントの絞り込みは困難を極める。したがって、ゲノム情報の違い（バリアント=遺伝子変異・多型）が、分子・細胞のレベル、ひいては表現型（疾患の発現や症状）においてどのような影響を及ぼすかといった情報（遺伝子型-表現型相関）から個々のバリアントの病的意義の大きさを定量的に見積もる統計学的手法の開発が求められる。他方、ゲノムや分子情報、それ自身を目的別に関連付け（実験）した生物学・ゲノム医科学的知見が公共データベース（DB）として溢れている。この中には、多くの有益な知見が隠れているにもかかわらず、蓄積された膨大な情報の中から手探りで有益な情報を見つけることは難しい。

２．研究の目的

本申請課題は、以下の２点を目的とした。

１．データマイニングによる遺伝性疾患（モデル疾患）の遺伝子型-表現型相関の明確化

２．データマイニングにより見出した各種遺伝子型-表現型相関を予測因子として、新規バリアントにおける病的意義の大きさを見積もる表現型予測モデルを提案

３．研究の方法

本提案手法によるデータマイニングによる遺伝子型-表現型予測因子の同定と表現型予測モデルの構築手法の概念図を Fig1 に示した。モデル疾患を対象にデータマイニングによる遺伝子型-表現型相関の解析（パターンの抽出）を行い、得られたパターンを予測因子として表現型予測モデルを構築した。表現型予測モデル構築に適用する機械学習法は、サポートベクタマシン（SVMs)、ランダムフォレスト、ニューラルネットワークなどの中から抽出されたパターンに対して長所を最も活かせるものを評価・選定することとした。構築した表現型予測モデルの分類精度は、sensitivity, specificity などにより多角的に評価し、最も分類精度が高くなるよう各種確率モデルのパラメータを設定した。この際、モデル疾患データセットを用いて、クロスバリデーション法（10-fold cross-validation）により評価を行った。モデル疾患データセットとしては、先行研究において扱ってきた *SCN1A* 関連てんかん性脳症とその類縁疾患で報告されている *SCN1A* ミスセンス変異を対象とした（先行研究：研究活動スタート支援（22890079)、若手研究（B）（24790336）以降に新たに報告された変異を加え再構築)。

４．研究成果

モデル疾患を対象にデータマイニングによる遺伝子型-表現型相関の解析（パターンの抽出）を行ったところ、Table1 に示す予測因子が得られた。次に得られた予測因子と種々の機械学習法を組合わせて、表現型予測モデルを構築した。構築した表現型予測モデルは、sensitivity, specificity 等の予測精度指標をクロスバリデーション法により算出・評価した。その結果、予測因子としてミスセンス変異によるアミノ酸の物理化学的特性変化（極性 P、疎水性 HP、等電荷点 IE）と変異の局在を用い、SVMｓにより機械学習した表現型予測モデルが、accuracy 0.88, sensitivity 0.94, specificity 0.74 と最も高い予測精度を示した。今後は、本提案手法により構築した表現型予測モデルの他疾患への応用を検討する。

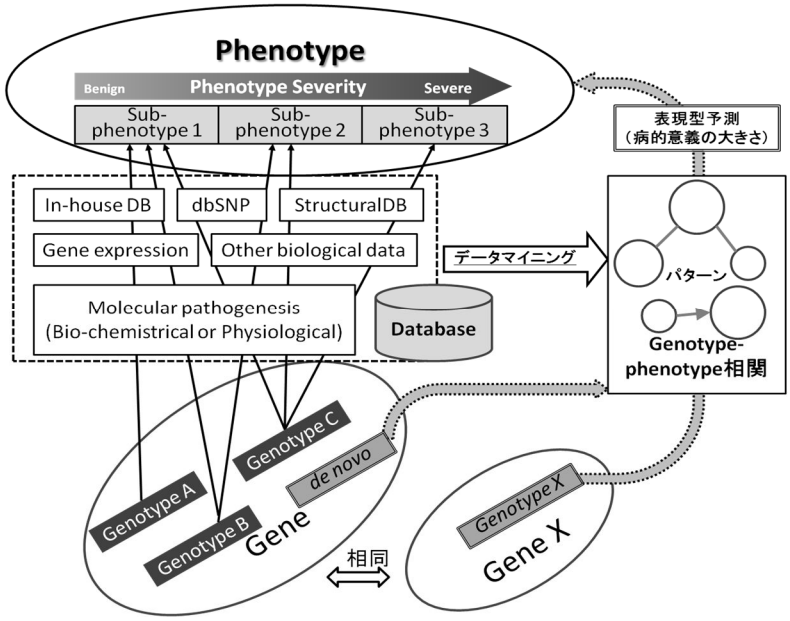

Fig1. **表現型予測モデル構築の概念図。**
genotype と phenotype（sub-phenotype）**の間には、直接的な相関が見られなくとも、データマイニングを用いることにより、関連する生物学・ゲノム医科学** DB **から遺伝子型-表現型相関を見出し、表現型予測モデルを構築することが可能となる。概念図の例では、各種** DB **に埋もれた既知の** genotype A,B,C **と** sub-phenotype **の相関関係をデータマイニングにより見出し、表現型予測モデルを構築。候補バリアントである同一遺伝子の** *de novo* **変異や相同な** gene X **の** genotype X **に表現型予測モデルを適用することで、病的意義の大きさを見積もる。**

Table1. **モデル疾患を対象とした遺伝子型-表現型解析により得られた予測因子**

Predictors	Effect site	Description（Genotype-phenotype correlation）
Localization	Whole	The missense mutations in the pore regions were associated with SIREE phenotype than mutations in other regions.
Absolute value changes of IE	Pore	The absolute value change of IE were significant difference between SIREE phenotype and GEFS+ phenotype.
Value difference of HP	S1–S4	Compare with SIREE phenotype, the GEFS+ phenotype has significant low value difference of P.
Value difference of PR	S1–S4	Compare with SIREE phenotype, the GEFS+ phenotype has significant low value difference of PR.
Value difference of P	Whole S1-S4	Compare with SIREE phenotype, the GEFS+ phenotype has significant high value difference of P. Compare with SIREE phenotypes, the GEFS+ phenotype has significant high value difference of P.

IE:isoelectricpoint; HP:hydrophobicity; P:polarity; S: transmembrane segment.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 吉田秀一、兼子直	４．巻 35
２．論文標題 てんかんと遺伝子解析	５．発行年 2017年
３．雑誌名 Clinical Neuroscience	６．最初と最後の頁 788-791
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 S Yoshida, T Nishio
２．発表標題 A Datamining Approach for Genotype-phenotype Correlation of SCN1A-related Epilepsies Based on Physico-chemical Properties Changes
３．学会等名 The 56th Annual Meeting of The Biophysical Society of Japan
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	兼子　直 (Kaneko Sunao) (40106852)	湊病院・北東北てんかんセンター・センター長	
連携研究者	西尾　卓広 (Nishio Takuhiro) (90172626)	浜松医科大学・医学部・准教授 (13802)	