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　３ 年 １０ 月 ２１ 日現在

機関番号：３２６１２
研究種目：挑戦的研究（萌芽）
研究期間：2017～2018
課題番号：１７Ｋ１９９２７

研究課題名（和文）深層学習エンジンを用いた疾患予測システムの研究

研究課題名（英文）Investigation of the disease prediction system using the deep learning engine

研究代表者
満山　進（Mitsuyama, Susumu）

慶應義塾大学・医学部（信濃町）・特任助教

研究者番号：３０２９６７２７

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：深層学習プログラムを使用して、タンパク質ドメイン中の遺伝子変異と相関がある疾患の予測システムの構築を行った。その結果3-methylcrotonyl-CoA carboxylase deficiency、Cystic fibrosisなど6疾患について相関が見られた。この結果は、Webページ(http://cancerproview.info/disease/)で検索を行える。また、遺伝子パスウェイを用いた疾患関連遺伝子検索ツールの作成を行った。

研究成果の学術的意義や社会的意義
本研究により開発された深層学習を用いたプログラムを使用することにより、タンパク質ドメイン、遺伝子変異と疾患との相関性が6疾患について明らかになった。今回使用したデータベース以外からより多くの遺伝子変異データを収集し、コンピューターに学習させることでタンパク質ドメインと疾患の相関性の解析がさらに進展すると考えられる。今後このシステムは、病気の診断や治療、創薬などに応用されると考えられる。

研究成果の概要（英文）：We developed the disease prediction system to be associated with gene mutation in a protein domain for using deep learning program. As a result, correlation of disease was about 6 diseases including 3-methylcrotonyl-CoA carboxylase deficiency, Cysticfibrosis etc. We made Web page (http://cancerproview.info/disease/) for this result search. In addition, We developed the disease-related gene search tool using the gene pathway.

研究分野：分子生物学

キーワード：深層学習　予測　モデル化　疾患　遺伝子変異　タンパク質ドメイン　解析ツール　データベース

１．研究開始当初の背景

(1) ヒトゲノム解析計画(Human Genome Project)が、2003 年（引用文献①）に終わったが、その後､次世代シーケンサーが開発され､ヒトのゲノム解析のデータ量とスピードは格段に進歩した。その後、2008 年に 1000 人ゲノムプロジェクトが始まり 2012 年（引用文献②）に終了した。タンパク質については、ヒトゲノム解析計画の終了以前の 2002 年よりタンパク 3000 プロジェクトが開始され 2006 年に終了した。これらの遺伝子、タンパク質の解析プロジェクトによりヒトの遺伝子／タンパク質の情報は飛躍的に増え、さらにヒト遺伝子の一塩基多型（SNPs: Single Nucleotide Polymorphism）などを含めるとその情報量は膨大である。現在、HUGO（Human Genome Organisation）の遺伝子記号の検索サイトである HGNC（HUGO Gene Nomenclature Committee)には、タンパク質を産生する遺伝子が 2016 年 10 月の時点で 19,018 遺伝子ある。ヒトの遺伝子疾患のデータベース OMIM（Online Mendelian Inheritance in Man）には 2016 年 10 月現在 8,329 の遺伝子疾患に関する記述があり、ヒトの単因子疾患の遺伝子変異データベースである HGMD（The Human Gene Mutation database）には、疾患と関係がある遺伝子変異がある遺伝子は、cDNA の数で 7,709 登録されていた。一塩基多型の頻度と疾患や遺伝的形質などの関係を統計的に調べるゲノムワイド関連解析（GWAS: Genome-Wide Association Study）のデータベースである European Bioinformatics Institute（EMBL-EBI）の GWAS Catalog（引用文献③）など GWAS のデータベースも近年増加する傾向にあり、その情報量は飛躍的に増えてきている。

(2) 近年、人工知能（AI）の分野の発達が目覚ましく、東京大学医科学研究所の研究では IBM によって開発された「自然言語応答システム」Watson によって特殊な白血病患者の病名を診断することに成功したと報告された。このように人工知能の医学・生物学への応用は、実用段階に入っていると考えられる。人工知能と関連した研究の中でも機械学習は単一のニューラルネットワーク（Neural network）が利用されその後、研究の進歩により近年、その発展形である多層構造形成している深層学習（Deep learning）が開発され応用されるようになった。深層学習を使うことにより、精度の高い遺伝子／タンパク質と疾患の関連性が、予測できると考えられる。

(3) 研究代表者はこれまで、がん関連疾患遺伝子／タンパク質相互作用データベース *CancerProView*（引用文献④、科研費研究課題／領域番号 18590538 基盤研究（C）: H18-H19「癌関連遺伝子・タンパク質パスウェイの画像表示システム開発」研究代表)、単因子遺伝子疾患変異データベース *MutationView*（引用文献⑤、⑥）などの疾患関連遺伝子／タンパク質データベースの研究を行ってきた。*CancerProView* の機能として疾患関連タンパク質の機能性ドメインと疾患の対応表を作成し、疾患との関連が不明な遺伝子が産生するタンパク質の機能性ドメインを比較することによって疾患との関連性を表示することができるシステムの作成を行った（科研費研究課題／領域番号 24590704 基盤研究（C）: H24-H26「癌関連疾患予測システムの開発と疾患パスウェイの研究」研究代表者)。その研究過程において、同じタンパク質ドメインを持っていても違う疾患であるものが多数あることが判明した。これらの遺伝子がどのような疾患になるかを予測することで、病態の解明や創薬に役立つと考えられる。

２．研究の目的

(1) 遺伝子／タンパク質、タンパク質ドメイン、多型と疾患について深層学習を用い相関性解析を行うプログラムの開発を行い、病気の診断や創薬、治療に応用できるシステムの開発を目的としている。

(2) 本研究は、研究代表者が開発した *CancerProView* を拡張させることによりシステムの構築を行う。

３．研究の方法

(1) 疾患関連タンパク質機能ドメインデータの収集

ヒト遺伝子タンパク質アミノ酸残基配列を米国 NCBI（National Library of Medicine National Institutes of Health）のタンパク質データベースの中の Refseq（Reference Sequence Database）から 113,373 件のデータ取得を行った。得られた配列は、EMBL-EBI の InterProScan を用いてタンパク質機能ドメインの検索を行い延べ 275,265 のタンパク質機能ドメインが得られた。得られたデータは、Refseq の ID と対応させアミノ酸残基配列上の位置、cDNA 塩基配列上の位置、タンパク質機能ドメイン名、遺伝子記号の表を作成した。

(2) 疾患関連遺伝子の遺伝子変異データの収集

疾患関連遺伝子変異データは、HGMD から遺伝子変異データを入手した。データの取得は Web 上で行い、8,613 遺伝子についてデータを得られた。その後、HGMD のデータベースに記載のある Refseq ID と対応するように Refseq ID、アミノ酸配列上の位置、cDNA 上の位置、遺伝子記号の表を作成した。

(3) 一塩基遺伝子多型データ（SNPs）の収集

疾患関連 SNPs データの収集は、NCBI の一塩基多型データベース（dbSNP）の ID を取得する。その後、NCBI のサイトから SNPs データを取得した。387,733 件について取得することができた。元となる遺伝子／タンパク質から Refseq ID を取得し Refseq ID、アミノ酸配列上の位置、cDNA 上の位置、遺伝子記号の表を作成した。

(4) 疾患関連遺伝子／タンパク質相互作用データの収集
疾患遺伝子／タンパク質相互作用のデータは、研究代表者が作成した *CancerProView* のデータを用いるとともに KEGG（Kyoto Encyclopedia of Genes and Genomes）から 87 件と米国 National Cancer Institute の BioCarta から 5,517 件のデータを用いて、その系内の遺伝子について関係する疾患、遺伝子記号の表を作成した。さらに解析データとして使用するために *CancerProView* に収録されている NCBI のタンパク質ドメインデータの更新を行った。
(5) 深層学習を用いた解析システムの構築
深層学習を用いた解析プログラムと遺伝子疾患予測プログラムの開発のために購入した GPU（Graphics Processing Unit）サーバー上に、米国 Google が無償で供給している深層学習プログラム Tensorflow をインストールし、開発環境の構築を行った。
(6) 遺伝子変異に対応付けしたタンパク質ドメインデータの作成
HGMD のデータベースに記載のある Refseq ID と対応するように Refseq ID、アミノ酸配列上の位置、cDNA 上の位置、遺伝子記号の表を作成した。ドメイン名、遺伝子シンボル、疾患・表現型のユニークな組み合わせは、16,681 個であった。そのうち、変異数の値が 1 であるものは 10,397 であった。変異数が 1 である割合が非常に高かった。データのばらつきは、機械学習の精度に影響を及ぼすため、疾患とドメインとの対応が 1 つだけの場合は、該当する疾患とドメインを学習用データセット生成時に除外した。
(7) 深層学習による疾患予測システムの開発
(1)、(2)、(3)、(4)、(5)で作成した表を使用して遺伝子／タンパク質、タンパク質ドメイン、疾患の関連性について解析を行うプログラムの作成は、python を用いて行った。タンパク質ドメインの中の変異が一つのものを除いた 1,477 個のデータを用いて、深層学習による疾患予測プログラムの開発を行った。このプログラムは入力層、中間層、出力層からなる 3 層ニューラルネットワークを使用して(6)で作成したデータを用いて学習モデルの構築を行った。
(8)遺伝子パスウェイ検索ツールの作成
KEGG や BioCarta に対応した遺伝子パスウェイ上の関連遺伝子検索ツールの作成を行った。プログラムは、python 上の tensorflow のライブラリである pandas、numpy、xkrd を使用して作成を行った。
(9) 解析結果の検索システムの構築
(6)で解析され、遺伝子疾患予測された結果は、インターネット上で公表することが可能な Web 検索システムの構築を行った。システムの構築を行う前に *CancerProView* のシステムのバックアップの作製を行い、その後にシステムの構築を行った。検索システムの入力インターフェースには PHP を用いた。データベースには、一般的なリレーショナルデータベースである MySQL を使用した。

４．研究成果
(1)遺伝子変異データに対応付けしたタンパク質ドメインデータの作成
取得した米国 NCBI の Refseq ヒト遺伝子タンパク質アミノ酸残基配列のデータ 113,373 件について、EBI-EMBL の InterProScan を使用して検索を行った 275,265 件のタンパク質ドメインと疾患関連遺伝子変異データベース HGMD の 8,613 遺伝子の遺伝子変異を用い、タンパク質ドメインと遺伝子の領域について対応付けを行い、Missense、Small deletion、Small insertion、Small indel の各変異についての一覧の作成を行った。HGMD の中には NCBI の一塩基型データベース（dbSNP）の病原性遺伝子多型のデータを含んでいる。その結果、4,929 遺伝子、2,522 個のタンパク質ドメインと疾患の対応付けをすることができた。
(2)深層学習による疾患予測システムの開発
作成した深層学習を用いた疾患予測プログラムによりタンパク質ドメイン中の遺伝子変異と関連する確率の高い疾患の予測を行った。その結果、3-methylcrotonyl-CoA carboxylase deficiency、Adrenoleukodystrophy、Cystic fibrosis、Glutaric acidemia、Gyrate atrophy、Stargardt disease の 6 疾患について相関が見られた。
(3)遺伝子パスウェイ検索ツールの作成
このツールは、以下の形式によりコマンドラインで使用する。
kegg2gene.py 遺伝子記号
このコマンドを実行すると遺伝子パスウェイから関連性のある遺伝子の探索を行うことができる。結果は、図 1 のように関連が表示される。
このプログラムにより、関連する遺伝子を簡便に検索することができる。

**** GenePools *******
CDKN2A, ARF, CDK4I, CDKN2, CMM2, INK4, INK4A, MLM, MTS-1, MTS1, P14, P14ARF, P16, P16-INK4A, P16INK4, P16INK4A, P19, P19ARF, TP16
CDK4, CMM3, PSK-J3
RB1, OSRC, PPP1R130, RB, p105-Rb, pRb, pp110
E2F1, E2F-1, RBAP1, RBBP3, RBP3...
CDKN2C, INK4C, p18, p18-INK4C
CDKN1A, CAP20, CDKN1, CIP1, MDA-6, P21, SDI1, WAF1, p21CIP1
CDKN1B, CDKN4, KIP1, MEN1B, MEN4, P27KIP1

図 1. 関連遺伝子検索の結果 (例：TP53)

(4)インターネットからアクセス可能な Web 検索システムの開発
本研究の成果は、インターネットから Web ページ（URL: http://cancerproview.info/disease/）にアクセスし、遺伝子記号を入力することにより検索を行うことができる（図 2 参照）。
検索された結果は、遺伝子記号、タンパク質ドメイン、疾患名、相関係数の順で一覧表示される。

図 2.検索画面

(5)研究総括
本研究によりタンパク質ドメイン、遺伝子変異と疾患との相関性が 3-methylcrotonyl-CoA carboxylase deficiency、Adrenoleukodystrophy、Cystic fibrosis、Glutaricacidaemia、Gyrate atrophy、Stargardt disease の 6 疾患について明らかになった。研究成果を Web 検索システムにより公開することで、今後、病気の診断や治療、創薬などに応用されると考えられる。

＜引用文献＞
① Shimizu N., Mitsuyama S. *et al*., (Author list 中 9 番目の慶應チーム), *Nature*, 2004 **431**:931-945
② 1000 Genomes Project Consortium, *Nature*, 2012, **491**:56-65
③ Walter D et al, *Nucleic Acids Research*. 2014, **42** (Database issue): D1001-1006.
④ Mitsuyama S, Shimizu N, *Genomics* 2012, **100**(2):81-92
⑤ Mitsuyama S, Ohtsubo M, Minoshima S, Shimizu N, *Human Mutatation*, 2015, **36**(8): E2430-2440
⑥ Minoshima S,Mitsuyama S, Ohno S, Kawamura T, Shimizu N, *Nucleic Acids Research*. 2000, **28**(1):364-368

５．主な発表論文等
〔雑誌論文〕（計 0 件）

〔学会発表〕（計 4 件）（うち招待講演 0 件／うち国際学会 0 件）

1. 発表者名 満山　進
2. 発表表題 がん関連疾患遺伝子/タンパク質相互作用データベース *CancerProView* の新展開
3. 学会等名 第 77 回日本癌学会学術総会
4. 発表年 2018 年

1. 発表者名
満山　進

2. 発表表題
がん関連タンパク質/遺伝子相互作用データベース *CancerProView* の新展開

3. 学会等名
2017 年度生命科学系学会合同年次大会

4. 発表年
2017 年

1. 発表者名
満山　進

2. 発表表題
がん関連疾患遺伝子/タンパク質相互作用データベース *CancerProView*

3. 学会等名
トーゴーの日シンポジュウム

4. 発表年
2017 年

1. 発表者名
満山　進

2. 発表表題
CanceProView:がん関連疾患遺伝子/タンパク質相互作用データベース

3. 学会等名
第 76 回日本癌学会学術総会

4. 発表年
2017 年

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

CancerProView ホームページ（Predict Disease Search）
http://cancerproview.info/disease/

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	森　努 （MORI Tsutomu）		