令和2年度　修士論文

イジングマシンを用いた正確ロジスティック回帰

Exact Logistic Regression with Ising Machines

2021年2月5日提出

東京大学大学院新領域創成科学科

メディカル情報生命専攻

指導教員　津田 宏治 教授

佐藤　史歩

Shiho Sato

要旨

統計検定は、生物学の研究において不可欠であり、通常は、漸近近似に基づく手法が用いられている。しかし、サンプル数が少ない場合には、誤差が大きくなるという問題が生じるが、全列挙に基づく正確検定手法は計算量が大きすぎるため、従来の計算機では実行が困難である。一方、量子アニーラーを初めとするイジングマシンは、近年急速に発展しており、多くの応用が期待されている。本論文では、イジングマシンを用いたサンプリングにより、正確ロジスティック回帰を行う手法を提案する。本手法では、適切なハミルトニアンを定義し、イジングマシンを用いてその基底状態を求めることで、帰無分布のサンプリングを行う。古典計算機上でのシミュレーテッドアニーリング、及び、量子アニーラ D-Wave 2000Q を用いた量子アニーリングを用いて、提案法の実装を行った結果、全列挙に比較して、大幅に計算時間を削減しながら、正確な P 値の推定が可能であることがわかった。

目次

第 1 章　はじめに **3**

1.1　背景：ゲノムワイド関連解析とロジスティック回帰 3

1.2　背景：正確ロジステック回帰 (Exact logistic regression) 4

1.3　背景：イジングマシン 5

1.4　背景：量子アニーリング (QA) 7

1.5　主結果 8

1.6　本論文の構成 8

第 2 章　本論 **9**

2.1　イジングモデル 9

2.2　SA 10

2.3　QA 11

2.4　Exact logistic regression 12

2.5　イジングマシンを用いた Exact logistic regression 13

第 3 章　実験 **15**

3.1　実験環境 15

3.1.1　骨肉腫データ 15

3.2　実験結果 16

3.2.1　結果 1: チェーン強度とチェーンの破壊割合、チェーン強度とサンプル数 16

3.2.2　結果 2: 全列挙と SA, QA の比較 18

3.2.3　結果 3: SA と QA における計算時間の比較 19

3.2.4　考察 22

第 **4** 章　おわりに **24**

参考文献 **26**

第1章

はじめに

1.1　背景：ゲノムワイド関連解析とロジスティック回帰

2003 年にヒトゲノム計画が完了して以来、ゲノム配列の個人差と形質との関連についての研究が急速に進み、特に単一遺伝子疾患の原因遺伝子を探索する方法として positional cloning が多用されてきた。これは 1980 年代に開発された、多型マーカーを目印に、疾患原因遺伝子と多型マーカーが同じ染色体上にある状態（連鎖）を見つけ出し、疾患原因遺伝子の位置と配列を段階的に突き止めていく手法である [25]。ところが実際は単一あるいは数個の遺伝子の異常のみで説明の付く病気はわずかで、大半が複数の遺伝子に少しずつ影響されることで発症する多因子疾患だと判明してきたため、ゲノム全体を巨視的に見渡す手法として、ゲノムワイド関連解析（Genome Wide Association Study; GWAS）が用いられるようになった。

GWAS は特定の疾患の非血縁の患者集団と非血縁の健常対照集団との間で、有意に関連する遺伝マーカーをゲノム全域にわたって網羅的に検索する手法である [25]。遺伝マーカーには遺伝的多様性を表す一塩基多型 (single nucleotide polymorphism; SNP) が使われている。2005 年に加齢黄斑変性を対象とした GWAS が行われて以来、現在までに 4800 を超える論文、24 万を超える形質との関連が発見されている [9]。統計モデルとしてロジスティック回帰を適用することが多い [8]。

ロジスティック回帰は、関数を用いて 0-1 の間となるよう共変量とパラメータの線形結合の和を変換させ、得られる値を確率と捉えることで、応答変数が二値であるデータに適用をもたせた回帰モデルである。例えば、ある事件が発生するか否かをロジスティッ

ク回帰に当てはめると、得られた値はその事件が発生する確率となる。そして分析した結果、確率が 0.5 より大きいとその事件が発生すると予測し、0.5 を下回るとその事件が発生しないと予測する。

ロジスティック回帰における統計解析は最尤推定と尤度比検定を用いて行う。最尤推定とはパラメータを変数、説明変数を定数とした際に実測値が得られる確率を尤度関数というが、その最大確率を知る手法である。最大確率をもたらすパラメータを最尤推定量という。尤度比検定とは最尤推定量を入れた尤度関数を用いて、帰無仮説の棄却を行う統計検定である。まず、帰無仮説に対する最尤推定量、対立仮説に対する最尤推定量それぞれを求める。次にそれぞれを尤度関数に代入し、その比をとる。ここで得られた比が、漸近的には自由度 $(H_0 - H_1)$ の χ 二乗分布に従うことを利用したのが尤度比検定である [28]。

これらの統計解析はいくつか問題を抱えている。まず尤度比検定はサンプル数が大きいことを想定した検定であるため、サンプルが小さい場合に不正確になってしまうという問題がある。また、独立性について検定するため分割表を使用する際、疎であると最尤推定ができない場合も多い。偏りの大きいデータに対しても正確に検定を行うことが難しい [1, 16, 19, 20, 24]。それらの問題を回避する方法の一つとして正確ロジスティック回帰が提案されている。

1.2 背景：正確ロジステック回帰 (Exact logistic regression)

正確ロジステック回帰 (Exact logistic regression) は Cox によって 70 年代に提案された統計的仮説検定で、関心のある回帰係数パラメータに対応した十分統計量の正確なパーミュテーション分布を、それ以外の回帰係数パラメータに対応した十分統計量を固定（条件つけ）して推定する、正確な条件付き推定である [4]。この手法は、得られる推定値は漸近的な結果に依存しておらず、帰無仮説の元で検定統計量がその値となる確率である p 値や信頼区間に対する正確な数値が得られる。そのため、最尤推定を適用できない場合、すなわち通常のロジスティック回帰に対してサンプルサイズがとても小さい場合、データが疎あるいは同じ値をもつなど偏っている場合に Exact logistic regression は有効である [1, 19, 20, 24]。一方この手法は計算が NP 完全であるという重大な欠点をもつ。これま

で、Cox が発表してから 10 年以上経った 1984 年にはじめてアルゴリズムが提案されて以降 [24]、最近ではマルコフ連鎖モンテカルロ法を用いたアルゴリズムなども発表されているが [21]、多くの変数を持った大規模データに対し適用可能なアルゴリズムは未だに開発されていない。

1.3 背景：イジングマシン

イジングモデルは、社会に潜む課題を組合せ最適化問題として定式化することで、高速に解くことができる手法である。分野を問わず数多くの課題が適用可能であり、例えば創薬分野における分子類似性比較問題 [27] や金融におけるポートフォリオ最適化問題 [2] などが挙げられる。またイジングモデルを利用した、人工知能を支える機械学習アルゴリズム開発も行われている。よって、さらなる研究開発が進み、高速に大規模問題を解くことができるようになることが期待されている。

このような社会背景から、現在イジングモデルを搭載したあらゆるイジングマシンが開発されている。以下の表 1.1 にイジングマシンの具体例を示す [6, 10, 17, 26, 30–32]。大別すると、量子アニーリング (quantum annealing; QA) とシミュレーテッドアニーリング (simulated annealing; SA) がある。QA は高速で、より SA よりも厳密解に近い答えが得られるとされている一方、解くことのできる問題サイズが限定されている。SA は、量子アニーリングよりも大きな問題を解くことができ、特別な冷却装置不要で常温で安定動作が可能である。例えば文献 [10] のように各マシンを比較検討した論文もあるが、実際には両方式を正確に比較した研究が少ないの現状である。

表 1.1: 各メーカーのイジングマシンの仕様比較

(A) SA 方式イジングマシン

	富士通	日立	TOSHIBA	NEC
製品名	デジタルアニーラ	CMOS アニーリング	シュミレーテッド分岐マシン	SX-Aurora TSUBASA
ステータス	商品化	商品化	商品化	商品化
実装方式	ASIC	GPU/ASIC・FPGA	HW に非依存	ベクトル計算機
規模	8, 192	100, 000	100, 000	100, 000
結合の仕方	全結合	全結合/疎結合	全結合	全結合

(B) QA 方式イジングマシン

	D-wave	産総研
製品名	D-wave 2000Q	未公開
ステータス	商品化	研究開発中
実装方式	超電導回路	超電導回路
規模	2, 048	未公開
結合の仕方	隣接結合	未公開

(C) レーザー方式イジングマシン

	NTT
製品名	コヒーレントイジングマシン
ステータス	研究開発中
実装方式	レーザー + FPGA
規模	2,000
結合の仕方	全結合

1.4 背景：量子アニーリング (QA)

暗号解読のため第二次世界大戦の最中開発されたコンピュータは、真空管からトランジスタへと材料は変わったものの、ノイマン型と呼ばれる方式で現在まで開発されてきた。性能の改善は、CPU などの機能を作り出す集積回路に含まれるトランジスタの数を増やす（トランジスタの大きさを小さくする）ことで行われてきた。しかしこの従来の方法による性能改善が限界を迎えている。トランジスタの大きさをこれ以上小さくしようとすると原子以下になってしまい、今までの考え方が通用しなくなるからだ。これは一般にムーアの法則の限界として知られている。そこで、新素材開発、あるいは非ノイマン型のコンピュータ開発が進められている。

またコンピュータにおける消費電力量も問題となっている。2013 年時点でコンピュータを中心とする IT が消費する電力は年間 1500TWh にのぼり、世界の発電量の 1 割にもなっている。Google, Amazon といった巨大 IT 企業が冷却に必要な電力を減らすため比較的気温の低い地域にデータセンターを設置したり、再生エネルギーの導入に腐心していることから危機感が見て取れる。

そのような背景のもと、近年 QA が注目されている。QA は西森らによって 1998 年に開発された、量子効果を用いて組合せ最適化問題を解く特化型コンピュータである [15]。量子は原子あるいはそれ以下の大きさである物質やエネルギー単位の総称であり、古典物理学ではなく量子力学に従う。よって量子の重ね合わせや量子干渉といった性質を用いることで、高速計算が可能となる。

さらに消費電力も少ない。超伝導量子ビットを使った QA は、CPU とメモリを合わせたチップの冷却用以外にほとんど電力を使わない上、極低温に冷却する部分はシステムが大きくなっても小さいままであるため、よりずっと大きなシステムになっても消費電力は基本的には変わらないからだ。例えば QA の消費電力は 20kW である一方、スーパーコンピューター京は 12MW である。能力が発揮される分野が異なるものの、600 倍も節電が実現されている。ちなみに、日本の一般家庭では年間 400W 消費するため、京は 3 万軒分の消費電力である。

また、量子を用いるコンピューティングとしてゲート式という手法もあるが、QA はゲート式と比較してシステムの安定性に優れているため、搭載ビット数の多い実機が既

に商用マシンとして開発され、比較的容易に利用できる利点がある [29]。

1.5 主結果

本研究では QA を用いて Exact logistic regression を現実的な時間内に行う手法を提案し、実際に量子アニーラの実機を用いて実験を行なった。Exact logistic regression の P 値計算におけるサンプリングを、イジングモデルの基底状態サンプリングとして定式化した。基底状態サンプリングは、古典計算機における SA、及び QA を用いて実装した。各手法と比較検討した結果、SA、QA とも、全列挙と比べて大幅に計算時間を削減しつつ、Exact logistic regression を行うことが可能だとわかった。

1.6 本論文の構成

2 章では、以降の章で必要な用語と、イジングマシンを用いた Exact logistic regression を説明する。3 章では、実データを用いて、各手法を用いた P 値計算における、計算時間と精度の比較を行う。4 章では、 本論文をまとめる。

第2章

本論

2.1 イジングモデル

イジングモデルは、統計数学において強磁性体の振る舞いを数理的に考察するために導入された。上向き、または下向きのスピンが構成され、隣接するスピン間の相互作用及び外部から与えれた磁場の力によって状態が変化し、最終的には系全体のエネルギーが最小となることが期待される。与えられた格子上の格子点 i に磁性の根源であるスピンを意味する二値変数 $s_i \in \{\pm 1\}$ を定義する。強磁性体では隣接するスピン変数は相互作用の結果、互いに同じ値を取ろうとする傾向があると考えられる。この性質を示したハミルトニアンが以下の式 (2.1) である。

$$H_{ising}(s) = \sum_{i=1}^{N} h_i s_i + \sum_i \sum_{i<j} J_{i,j} s_i s_j \tag{2.1}$$

ここで $J > 0$ は相互作用の強さ、h は外部磁場、(ij) は隣接する格子点の対をそれぞれ表している。

また数学的に同じ意味であるモデルとして二次制約なし二値最適化 (Quadratic unconstrained binary optimization; QUBO) がある。こちらは、二値変数がバイナリとなっており、以下の式 (2.2) で表される。

$$H_{qubo}(x) = \sum_{i=1}^{N} a_i x_i + \sum_i \sum_{i<j} b_{i,j} x_i x_j + c \tag{2.2}$$

コンピューターサイエンスでは 0 ,1 の二値変数を扱うことのほうが多く、イジングモデルよりも QUBO が利用される。また QUBO は通常、a_i を非対角成分に、b_{ij} を対角成分においた上三角行列で表す。

イジングモデルからQUBOへの変換、またその逆への変換も可能である。具体手的には、イジングモデルにおける二値変数を

$$x_i \mapsto \quad \frac{s_i + 1}{2} \tag{2.3}$$

とすることでQUBOへ変換することが、QUBOにおける二値変数を

$$s_i \mapsto \quad 2x_i - 1 \tag{2.4}$$

とすることでイジングモデルへと変換を行うことができる。

イジング及びQUBOはどちらも二値二次モデル (binary quadratic model; BQM) に分類され、これらに定式化される問題は変数が増えるにつれて一般に解くことが難しくなり、基底状態を得るためにかかる時間が増加する。この性質をNP困難であるという [3]。

2.2 SA

SAは金属の焼きなましを模したモデルであり、熱揺らぎを用いて状態遷移を起こす。初期状態を高温とすることで活発な状態遷移を起こし、次第に温度を下げることで徐々に状態遷移を緩くし、最終的にグローバルな基底状態が得ることを期待する。動作原理からの観点では、マルコフ連鎖モンテカルロ法 (Markov chain Monte Carlo methods; MCMC) をアニーリングに適用させた手法である。MCMCは一つ前の状態にのみ依存して次の状態を作成することで、高次元・多変量の確率分布からサンプリングを行う手法である。SAでは、ある温度でMCMCを実行した後、温度を少し下げて同じ要領でMCMCを実行し、さらに少し温度を下げて実行するという仮定を繰り返すと、各温度での分布を順に追っていくことになる。最後に温度を0にすれば、温度0における分布、すなわち最低エネルギー状態（基底状態）が高い確率で実現する。SAをイジングモデルの基底状態を得るために利用する際は、エネルギーを関数と見たてた分布を考え、その分布から状態を生成するようなMCMCを実行をしつつ温度を下げていく [29]。

2.3 QA

QA はイジングモデルの基底状態を、量子効果を使って求める技術である。なんらかの課題を QA を用いて解くには、まずは課題を組み合わせ最適化問題としてモデル化し、イジングモデルあるいは QUBO に定式化したのち、量子アニーラに投入することで行うことが実施が可能となる。

QA の動作原理を以下の式 (2.5) に示す。

$$H_{qa}(t) = A(t)H_c + B(t)H_q \tag{2.5}$$
$$0 \leq t \leq \tau$$

$A(t)$ は単調増加関数を、$B(t)$ は単調減少関数を表す。

H_c は QA のイジングモデルを表す。

$$H_c = \sum_{i<j} J_{ij}\sigma_i^z\sigma_j^z + \sum_{i=1}^{N} h_i\sigma_i^z \tag{2.6}$$
$$\sigma_i^z \in \{\pm 1\}$$

σ_i^z はパウリ行列の z 成分であり、z 軸方向のスピンの大きさを表している。J_{ij} はスピン間の相互作用を、h_i は局所磁場を調整するパラメータを表している。

一方 H_q は量子効果を表す。

$$H_q = -\Gamma \sum_i \sigma_i^x \tag{2.7}$$

σ_i^x はパウリ行列の x 成分であり、スピンを反転させる作用を持つ。z 方向を縦とすると x 方向は横であるため、$B(t)H_q$ は横磁場項とも呼ばれる。

QA でははじめ、 横磁場項を大きくし、 各スピンの状態を不確定、 つまり ± 1 を等確率で重ね合わせた状態に設定する。そして、 量子力学の核であるシュレディンガー方程式に従って自律的に変化していくことで、量子効果が徐々に小さくなり、H_c における相互作用 J_{ij} や局所磁場 h_i の影響が大きくなっていく。すると各スピンの状態が ± 1 のどちらかに確定した状態が自律的に選ばれ、 それがイジングモデルの基底状態となる。つまり、 解きたい問題を ± 1 の値をとる変数を持つ式 (2.6) に落とし込み、 QA を行うと、その問題における目的関数 H_c が最低エネルギーをとるような σ が確率的に求まる [29]。

カナダの D-Wave Systems Inc. は QA の動作原理が実装されたハードウェアを世界で初めて開発し、現在ではクラウド経由で利用できる D-Wave2000Q が提供されている [14]。現在 D-Wave2000Q で扱うことのできる最大スピン数は 2048 個である。この D-Wave2000Q で採用されているグラフ構造はキメラグラフと呼ばれ、任意の $i, j (i < j)$ の組に対して $J_{i,j}$ を設定できるわけではない。そのためには、 複数のスピンを仮想的な 1 つのスピンとみなすことで、 全結合グラフを再現する必要がある。全結合グラフを作る際に扱えるスピン数は、 最大で 64 となる。エネルギーを最小にする基底状態が複数ある場合には、量子アニーラを用いて、基底状態のサンプリングを行うことができる [18]。本研究では、この性質を生かし、統計検定への応用を行う。

2.4 Exact logistic regression

Exact logistic regression とは、検定したい回帰係数パラメータの有意性を正確に推定する手法である。

まず以下に示すデータを与える。

- n 個の独立した標本データ
- それぞれのデータに対する共変量を以下のように定義する。
  - $x_{ij} \in \{0,1\}$, $\mathbf{X}_j = [x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}]$, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_m]$
  - 検定したい目的説明変数 $\mathbf{X}_k$(第 i 標本データにおいては x_{ik})
- それぞれのデータの応答変数は以下のように定義する
  - $y_i \in \{0,1\}$, $\mathbf{Y} = [y_1, \ldots, y_i, \ldots, y_n]$

以上のデータが与えられた際、共変量に対するロジスティック回帰モデルは、以下の式 (2.8) のように表される。

$$\log \quad \frac{p_i}{1-p_i} = \gamma + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} \tag{2.8}$$

$$p_i = \frac{1}{1 + \exp(\gamma + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im})}$$

このとき $x_{i1}, \ldots, x_{im}$ は第 i 標本における m 個の共変量、$\beta_1, \ldots, \beta_m$ は m 個の回帰係数パラメータ、p_i は上記の共変量と回帰係数パラメータが与えられた際に $y_i = 1$ となる確率を表している。

Exact logistic regression では、パラメータを関心のある目的変数の回帰係数パラメータと、それ以外である局外パラメータに分ける。そして局外パラメータの十分統計量の実現値で条件をつけて目的パラメータの推測を行う。

Exact logistic regression では十分統計量として以下を使用する。

$$t_0 = \sum_{i=1}^{n} y_i \tag{2.9}$$

$$t_j = \sum_{i=1}^{n} x_{ij} y_i \tag{2.10}$$

そして t_k 以外に対して以下のように条件付けをし、条件を満たす $\mathbf{Y}$ を全列挙する。$\hat{t}_j$ は元々のデータセットにおける $\mathbf{Y}$ に対して計算した十分統計量を表している。

$$\mathbf{Y} = \{y | t_0 = \hat{t}_0, t_j = \hat{t}_j, \ j \neq k\} \tag{2.11}$$

式 (2.11) を満たすすべての $\mathbf{Y}$ サンプルに対し t_k を計算し、その値以下の割合を p 値とする。

$$P(t_k \leq \hat{t}_k | t_0 = \hat{t}_0, t_j = \hat{t}_j) = \frac{\sum_{t'_k \leq \hat{t}_k} c(t_k = t'_k, t_j = \hat{t}_j)}{\sum_u c(t_k = u, t_j = \hat{t}_j)} \qquad j \neq k \tag{2.12}$$

ここで、$c(\alpha = \hat{\alpha}, \beta = \hat{\beta})$ は $\alpha = \hat{\alpha}, \beta = \hat{\beta}$ を満たす Y の個数を表している。以上が Exact logistic regression の手順である。

2.5 イジングマシンを用いた Exact logistic regression

本研究では、条件を満たす $\mathbf{Y}$ を得る過程にて全列挙を行わず、イジングマシンによるサンプリングにより評価する。イジングマシンにおいては、QUBO を最小化する基底状態（最適解）を探索する。基底状態が複数あるときは、文献 [23] でも示されているように、その中からサンプリングを行うことができる。Exact logistic regression をイジングマ

シンを用いて行うには、式 (2.11) を満たす解が基底状態になるようなハミルトニアンを設計する必要がある。ここでは以下のように設定した。

$$H = (t_0 - \hat{t}_0)^2 + \sum_{j \neq k} (t_j - \hat{t}_j)^2 \tag{2.13}$$

QA を用いた Exact logistic regression の手順は、以下の通りである。まず解きたい問題を QUBO 形式にし、量子アニーラへ入力する。量子アニーラにおいて複数回の基底状態探索を繰り返すことで、複数の基底状態をサンプリングすることができる。しかしながらアニーリングマシンでは、得られる状態が基底状態ではなく、励起状態（すなわち，式 (2.11) が満たされない解）が得られる場合がある。そのため、得られた解が望みの解になっているかは選別する必要がある。

Exact logistic regression では、検定したい特徴量と元のデータにおける応答変数との内積値 t_k がどれほど有意かを知ることによる統計検定であるため、得られたサンプルから、検定したい特徴量とサンプルとの内積値が $\hat{t}_k$ 以下となるサンプルの割合 p を計算する。

もし例えば有意水準を $\alpha = 0.05$ と定めていたら、$p < 0.05$ となるとき元のデータが有意である（対立仮説を棄却しない）とわかり、そうでない場合は元のデータは有意でない（帰無仮説を棄却しない）とわかる。

以上が量子アニーリングを用いた Exact logistic regression である。

第3章

実験

本章ではイジングマシンとして SA または QA を用いて Exact logistic regression を行い、その性能を全列挙手法を用いた Exact logistic regression と比較し、評価を行った。

3.1 実験環境

実験用のデータには、骨肉腫データを用いた [13]。実験用のプログラムは Python3.7.0 を用いて実装した。全列挙手法として 4ti2 ソフトウェア (https://4ti2.github.io/) に含まれる zsolve を利用した。SA として D-Wave Systems Inc. が提供する、イジングや QUBO 問題を解く Python 製パッケージ Ocean(https://docs.ocean.dwavesys.com) に含まれる neal の SimulatedAnnealingSampler を利用した。QA として D-Wave Systems Inc. の D-Wave2000Q を、クラウドサービスである Leap(https://cloud.dwavesys.com) を介して利用した。

3.1.1 骨肉腫データ

骨肉腫データは 46 標本を含む単変量解析結果のデータである [13]。骨肉腫データのうち特徴量は 3 つ使用し、スケーラビリティを評価するため、データ点を 20 点、25 点、30 点、35 点、40 点にダウンサンプリングしたものを使用した。各データ点数に関して、５種類のデータセットを用意した。以降、作成した 25 個のデータセットを骨肉腫データセットと呼ぶことにする。説明変数は文献 [13] で実施された多変量解析にて有意性が認められた共変量であり、１つ含まれている。説明変数はリンパ球浸潤の有無 (LI。有＝ 1、無＝ 0) を、それ以外の共変量は性別 (SEX。女性=1、男性=0)、オステオイド（類骨）に

なんらかの病理が観察されるか否か (AOP。有=1、無=0) である。応答変数も二値変数であり、無再発生存が 3 年以上だったか否か（DFI3。以上＝ 1、未満＝ 0）を示している。表 3.1 に詳細を記載する。

表 3.1: 骨肉腫データ

LI	SEX	AOP	DFI3=1 となるデータの割合
0	0	0	3/3 (100%)
0	0	1	2/2 (100%)
0	1	0	4/4(100%)
0	1	1	1/1(100%)
1	0	0	5/5(100%)
1	0	1	3/5(100%)
1	1	0	5/9(100%)
1	1	1	6/17(100%)

3.2 実験結果

以下に実験結果を示す。

3.2.1 結果 1: チェーン強度とチェーンの破壊割合、チェーン強度とサンプル数

D-Wave マシンはサンプリングを行う際、パラメータのひとつであるチェーン強度を調節することができる。チェーンとは、１つの論理ビットを表すために複数の物理量子ビットをつなぐものを指し、論理ビットをワーキンググラフ上に定義された物理量子ビットへと埋め込む際に必要に応じて用いる [6]。そこで D-wave を用いたサンプリングを実行するにあたり、事前にデータ点ごとに適切なチェーン強度を探索した。適切さは、チェーンの破壊割合とサンプル数によって判断した。

各骨肉腫データセットを用いて、チェーン強度を 10, 15, 20, 25, 30 ,35 と変化させサンプリングを実行した。結果を図 3.1 と、図 3.2 に示す。各々 y 軸にデータを、x 軸にチェー

ン強度を配置した。図 3.1 は誤差付き平均を、図 3.2 は分散が大きかったため平均のみを表している。

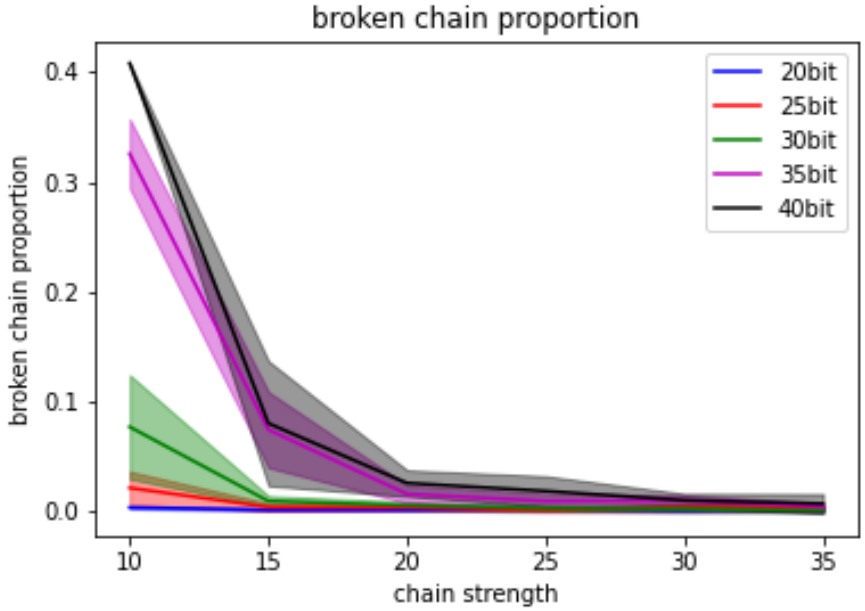

図 3.1: チェーン強度とチェーンの破壊割合

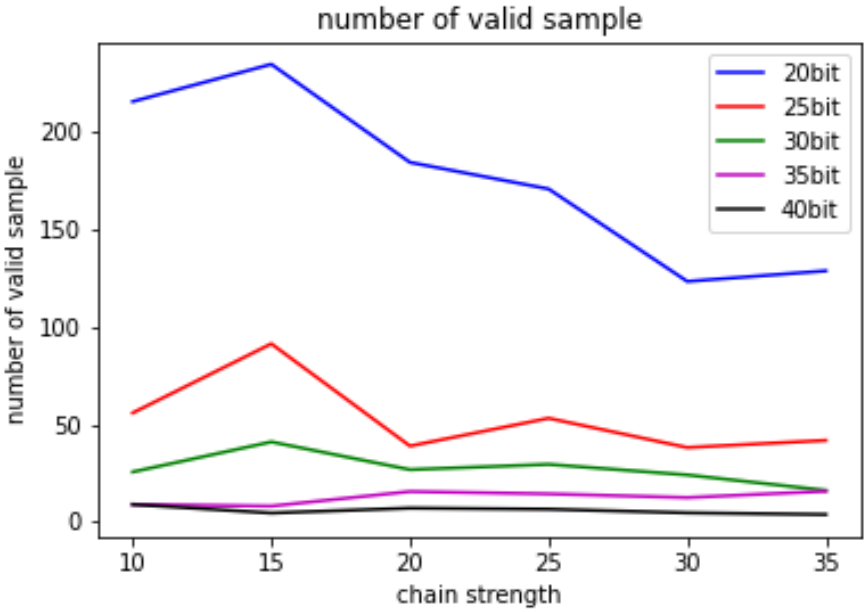

図 3.2: チェーン強度とサンプル数

図 3.1 より、データサイズが大きいほど弱いチェーン強度ではチェーンが壊れやすいこ

とがわかる。また図 3.2 から、データサイズが小さいほどサンプル数が多くなるチェーン強度が小さいことがわかる。以上の実験から 20-40 点のデータに対する最適チェーン強度を決定し、以下の表 3.2 にそれらの結果をまとめた。

表 3.2: データサイズと最適チェーン強度

	データ点				
	20	25	30	35	40
最適チェーン強度	15	15	15	20	20

3.2.2 結果 2: 全列挙と SA, QA の比較

3.2.1 より QA における各データ点に対するチェーン強度を決定したため、以降では Exact logistic regression におけるサンプリングを実行する。SA と QA に関して、各骨肉腫データセットから 10000 個のサンプルを生成する実験を行った。両手法とも、全てのサンプルが基底状態に達するわけではないため、基底状態に達したサンプルを有効サンプルと呼ぶことにする。全列挙との比較を目的としているため、全列挙が実施できた 30 点までの実験結果を以下の表 3.3、表 3.4、表 3.5 に示す。

表 3.3: 20bit での全列挙との比較

手法	計算時間	有効サンプル数	p 値誤差
全列挙	1.22	4.42×10^3	0
SA	4.25	2.60×10^3	4.21×10^{-4}
QA	2.40	2.04×10^2	6.89×10^{-2}

表 3.4: 25bit での全列挙との比較

手法	計算時間	有効サンプル数	p 値誤差
全列挙	5.84×10^2	2.79×10^4	0
SA	5.84	6.48×10^3	4.26×10^{-3}
QA	2.40	9.1×10^1	4.59×10^{-2}

表 3.5: 30bit での全列挙との比較

手法	計算時間	有効サンプル数	p 値誤差
全列挙	5.48×10^4	3.02×10^4	0
SA	7.60	9.21×10^3	6.11×10^{-3}
QA	2.40	8.2×10^1	6.26×10^{-2}

まず計算時間について、SA、QA どちらもとくにデータ点が大きくなるにつれて全列挙よりも大幅に短くなっており、SA では最大 7.2×10^3 倍、QA では最大 2.28×10^4 倍短くなった。

次に有効サンプル数について、SA は、どのデータ点においても全列挙には及ばないものの QA よりは多くの有効サンプル数を得られていた。

また、データ点数が増えるにつれて、ノイズの影響により基底解を見つけるのが困難になるため、QA の有効サンプル数は減少した。一方、ノイズの影響を受けない SA では、有効サンプル数は減少しなかった。

最後に全列挙との p 値誤差について、SA はどのデータ点においても最大でも 6×10^{-3} ほどの誤差に収まっていた。一方、QA は 6×10^{-2} ほど誤差が生じていた。

以上の結果から、どちらも Exact logistic regression における欠点である膨大な計算量を解決する手段としての可能性を示すことができた。しかし、現時点では SA のほうが有効であり、QA におけるサンプリングには改善の余地があることが分かった。

3.2.3 結果 3: SA と QA における計算時間の比較

SA と QA について、計算時間に注目をし比較を行った。両手法ともアニーリング箇所にかかる時間を計測した。とくに QA は `qpu_access_time` と呼ばれる時間を計測した [6]。

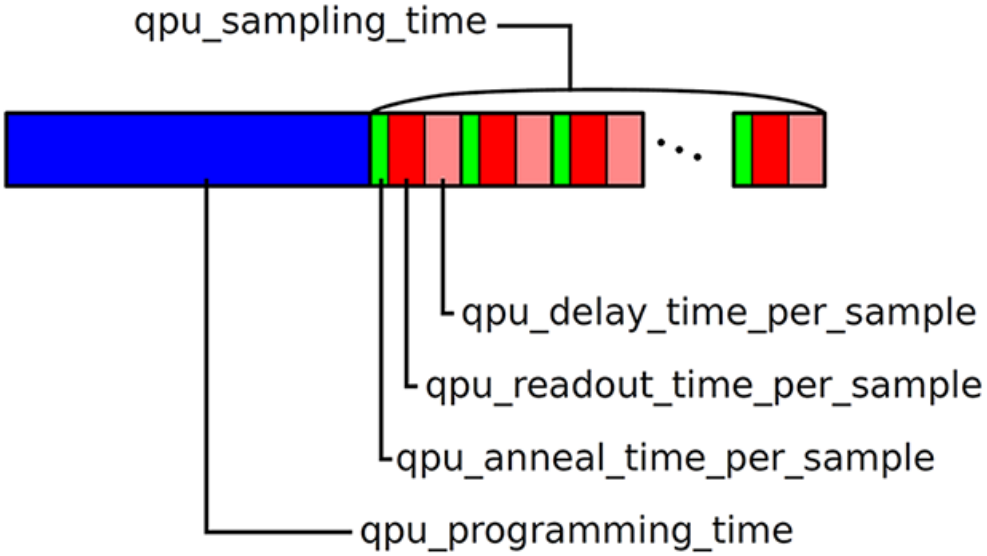

図 3.3: qpu access time の詳細

この時間は、qpu_programming_time と qpu_sampling_time に大別される。

qpu_programming_time には以下の工程が含まれている。

- h と J の値を QPU 上のデジタル・アナログ変換器 (DAC) へ伝達する。
- 電子回路が DAC をプログラムするための生の信号を生成し、それが冷凍機内へワイヤを介して送信される。
- DAC は量子ビットとカプラーに対し局所的に、静的な磁気制御信号を印加する。(以上をプログラミングサイクルという)
- QPU を冷却する (通常 $1ms$)。

qpu_sampling_time には以下の工程が含まれている。

- アニーリングを実行する。(annealing_time)
- 物理量子ビットのスピン状態を読み出して返す。
- QPU を冷却する。
- これをユーザーが指定した回数 (num_reads) 繰り返す。

また annealing_time は qpu_access_time に含まれ、任意で設定することが可能であるが、値を変えても有効サンプル数に変化が見られなかったので、初期設定時間を使用した。結果を図 3.4 に示す。両手法とも 40 点まで実行可能であり、有効サンプルを得ることができたため、40 点までに含まれる各 5 つの骨肉腫データの計測結果を、誤差範囲付き平均で表している。

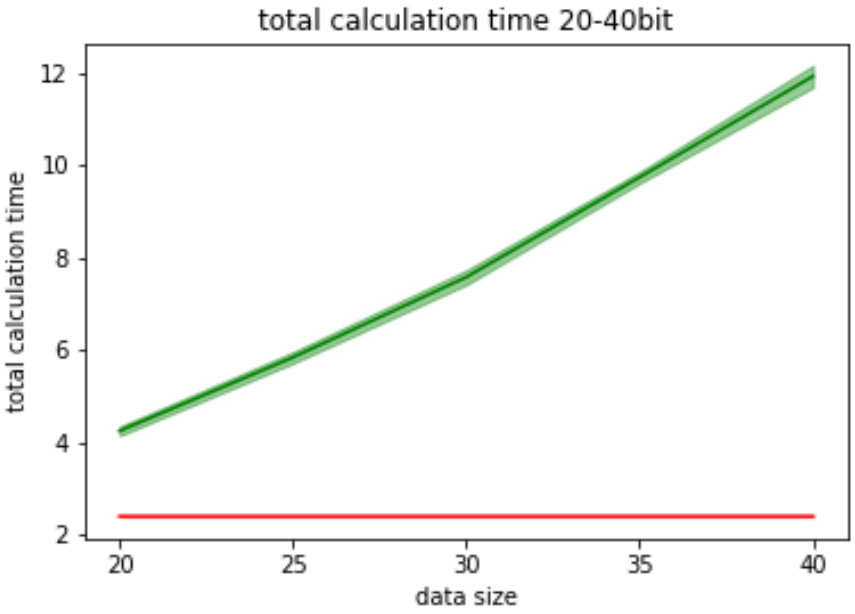

図 3.4: 計算時間の比較　(SA, QA)

結果図 3.4 から、SA はデータ点が増えるにつれて計算時間が増える一方、QA はデータ点が増えても計算時間が増えないことがわかる。

一般に、問題設定が同じである場合、データ点が増えれば増えるほど問題を解く際に必要な時間は増加する。そのため古典コンピューター上で解いている SA はたしかにデータ点が増加するにつれて時間が増加することが確認できる。一方量子性を用いた量子アニーリングは、物理デバイスを用いているので、デバイスの扱える範囲であれば、問題のデータ点が増えても、サンプリングにかかる計算時間は一定である。しかし、3.2.2 で示したように、QA は SA と比較して、得られた有効サンプル数がどのデータ点においても少ないという結果となった。これは、恐らく、以下に述べる統合制御エラーや、コヒーレンス時間の短さ、埋め込みが関連していると考えられる。

3.2.4 考察

3.2.2、3.2.3 から、QA は SA に比べて、基底解を得られる割合が少ないという結果となった。以下考えられる原因を挙げる。

まず、この問題が QA に不向きな問題であった可能性がある。現状では、QA が SA よりも優っているという研究もあれば [7]、その反例もある [22]。本研究は、最適解となる基底状態エネルギーが一つと決まっており、最適解に近い解である良解を解として含むことはできない問題、かつ QA をサンプリングとして利用した問題であった。しかし一般的には最適化問題として利用されることが多く、サンプリングとしての利用例は少ない。以降発表されであろう、基底状態エネルギーが既知かつ唯一であり、サンプリングとして利用している研究と比較を行いたい。

次に、統合制御エラー (ICE) が考えられる [6]。D-wave システムを利用して QA を行うと、たとえユーザーが $\{h, J\}$ を指定しても、D-wave QPU にこれらの値を実装すると、実際は以下の式 (3.1) のようにわずかに誤差を含む問題を解く。

$$H_{ice}(s) = \sum_{i=1}^{N}(h_i + \delta h_i)s_i + \sum_i \sum_{i<j}(J_{i,j} + \delta J_{i,j})s_i s_j \tag{3.1}$$

この原因は総称して ICE と呼ばれている。ICE の中には、接続されていない量子ビットに対してもあたかも接続しているかのような効果をもってしまう現象、(h, J) の大きさによるものなどが含まれる。これらの要因が合わさって、本来解きたい問題を正確にハードウェアに埋め込むことができず、厳密解を得ることができなかったと考えられる。この対処法として、スピン反転変換という方法がある。この方法は、いくつかの論理スピンの向きを変えることで、解く問題の性質を変えずに、係数の誤差を平均化して軽減する方法である。一つの論理スピンの向きを変更することで、

$$\begin{aligned} s \to s' &= -1 \\ s_p \to s'_p &= -s_p \\ h_p \to h'_p &= -s_p \\ J_{i,j} \to J'_{i,j} &= -J_{i,j} \text{ (for either } i = p \text{ or } j = p) \end{aligned}$$

で表されるゲージ変換が実現される。D-wave では変換する論理スピンの数を指定することができる。実際に本研究にてスピン変換をデータ点ごとに、$0-20, 0-25, 0-30, 0-35, 0-40$ 個変更し実験をおこなったが、改善は見られなかった。

また、現状の量子アニーラにおいてアニーリング時間に対してコヒーレンス時間が短い。コヒーレンス時間とは量子状態が存在している時間であり、とくに今回の D-wave 2000Q のように超電導ニオブ磁束量子ビットを用いた量子アニーラは、$100ns$ と言われている [11]。コヒーレンス時間が短いと、トンネル効果など量子特有の性質をもつ時間が短くなるため量子計算に悪影響が生じる。そこで現在各社コヒーレンス時間を延ばす研究開発を行っている。例えば MIT と Northlop Grumman は、コヒーレンス時間の長いアルミニウム磁束量子ビットと３次元実装技術を利用した量子アニーラの開発を IARPA の支援のもとで進めている [12]。さらに、埋め込みによる問題の破綻も考えられる。3.2.1 のようにデータや問題難易度に合ったチェーン強度をユーザーが任意で設定することが可能だが、重み $\{h, J\}$ が QPU 制御システムの精度限界を超えて圧縮されてしまう危険性がある。そして過剰に圧縮されることで、量子ハードウェアがマッピングされた個々の重み同士を区別することは困難になり、解の品質が低下する可能性があるという。また埋め込みにより生じる長いチェーンは、問題指定におけるエラーを増加させ、論理スピンの忠実度が低下することが知られている [5]。よって、今後、今回使用した D-wave のようなチェーンを利用する隣接結合デバイスではなく、全結合デバイスを利用した実験も行いたいと考えている。

第4章

おわりに

本論文では Exact logistic regression におけるサンプルを得る工程を、イジングマシンを用いて行う方法を提案した。そして実際に、 SA、または QA の実機である D-Wave2000Q を用いてサンプリングを行い、 全列挙と比較を行った結果を示した。結果から、SA または QA では、全列挙の手法では計算不可能なサイズのデータに対しても Exact logistic regression を行うことができることがわかった。また QA は SA ほど結果が振るわなかったが、今後 QA デバイスの改善によって、より良い性能が期待できる可能性がある。今後の展望として、 選択的推論への適用や、他のイジングマシンへの適用が挙げられる。

謝辞

終始熱心なご指導を頂いた津田 宏治教授、田村 亮講師に感謝の意を表します。北井孝紀さんをはじめ、津田研究室の皆様から多くの刺激と示唆を得ることができました。感謝の意を表します。本当にありがとうございました。

参考文献

[1] Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, 82(400):1110–1117, 1987.

[2] Clemens Adolphs Arman Zaribafiyan Alipour, Elham and Maxwell Rounds. Quantumin-spired hierarchical risk parity. http://1qbit.com/files/white-papers/1QBit-White-Paper-—Quantum-Inspired-Hierarchical-Risk-Parity.pdf. 2021-01-24 参照.

[3] Francisco Barahona. On the computational complexity of ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

[4] DR Cox. Analysis of binary data london: Methuen &co, 1970.

[5] D-wave. D-wave advantage system an overview. https://www.dwavesys.com/sites/default/files/14-1049A-A_The_D-Wave_Advantage_System_An_Overview_0.pdf. 2021-01-10 参照.

[6] D-wave. D-wave system documentation. https://docs.dwavesys.com/docs/latest/. 2021-01-10 参照.

[7] Vasil S Denchev, Sergio Boixo, Sergei V Isakov, Nan Ding, Ryan Babbush, Vadim Smelyanskiy, John Martinis, and Hartmut Neven. What is the computational value of finite-range tunneling? *Physical Review X*, 6(3):031015, 2016.

[8] David A Duverle, Shohei Kawasaki, Yoshiji Yamada, Jun Sakuma, and Koji Tsuda. Privacy-preserving statistical analysis by exact logistic regression. In *2015 IEEE Security and Privacy Workshops*, pages 7–16. IEEE, 2015.

[9] EMBL-EBI. Gwas catalog. https://www.ebi.ac.uk/gwas/home. 2021-01-10 参照.

[10] Ryan Hamerly, Takahiro Inagaki, Peter L McMahon, Davide Venturelli, Alireza Marandi, Tatsuhiro Onodera, Edwin Ng, Carsten Langrock, Kensuke Inaba, Toshimori Honjo, et al. Experimental investigation of performance differences between coherent ising machines and a quantum annealer. *Science advances*, 5(5):eaau0823, 2019.

[11] R Harris, J Johansson, AJ Berkley, MW Johnson, T Lanting, Siyuan Han, P Bunyk, E Ladizinsky, T Oh, I Perminov, et al. Experimental demonstration of a robust and scalable flux qubit. *Physical Review B*, 81(13):134510, 2010.

[12] IIARPA. Qeo. https://www.iarpa.gov/index.php/research-programs/qeo. 2021-01-12 参照.

[13] Norman Jaffe, John Knapp, Vincent P Chuang, Sidney Wallace, Alberto Ayala, John Murray, Ayten Cangir, Alexander Wang, and Robert S Benjamin. Osteosarcoma: Intra-arterial treatment of the primary tumor with cis-diammine-dichloroplatinum ii (cdp): Angiographic, pathologic, and pharmacologic studies. *Cancer*, 51(3):402–407, 1983.

[14] Mark W Johnson, Mohammad HS Amin, Suzanne Gildert, Trevor Lanting, Firas Hamze, Neil Dickson, R Harris, Andrew J Berkley, Jan Johansson, Paul Bunyk, et al. Quantum annealing with manufactured spins. *Nature*, 473(7346):194–198, 2011.

[15] Tadashi Kadowaki and Hidetoshi Nishimori. Quantum annealing in the transverse ising model. *Physical Review E*, 58(5):5355, 1998.

[16] Elizabeth N King and Thomas P Ryan. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, 56(3):163–170, 2002.

[17] Masaaki Maezawa, Go Fujii, Mutsuo Hidaka, Kentaro Imafuku, Katsuya Kikuchi, Hanpei Koike, Kazumasa Makise, Shuichi Nagasawa, Hiroshi Nakagawa, Masahiro Ukibe, et al. Toward practical-scale quantum annealing machine for prime factoring. *Journal of the Physical Society of Japan*, 88(6):061012, 2019.

[18] Salvatore Mandra, Zheng Zhu, and Helmut G Katzgraber. Exponentially biased ground-state sampling of quantum annealing machines with transverse-field driving hamiltonians. *Physical review letters*, 118(7):070502, 2017.

[19] Cyrus R Mehta and Nitin R Patel. Exact logistic regression: theory and examples. *Statistics in medicine*, 14(19):2143–2160, 1995.

[20] Cyrus R Mehta and Nitin R Patel. Exact inference for categorical data. *Encyclopedia of biostatistics*, 2:1411–1422, 1998.

[21] Cyrus R Mehta, Nitin R Patel, and Pralay Senchaudhuri. Efficient monte carlo methods for conditional logistic regression. *Journal of the American Statistical Association*, 95(449):99–108, 2000.

[22] Troels F Rønnow, Zhihui Wang, Joshua Job, Sergio Boixo, Sergei V Isakov, David Wecker, John M Martinis, Daniel A Lidar, and Matthias Troyer. Defining and detecting quantum speedup. *science*, 345(6195):420–424, 2014.

[23] Bhuvanesh Sundar, Roger Paredes, David T Damanik, Leonardo Duenas-Osorio, and Kaden RA Hazzard. A quantum algorithm to count weighted ground states of classical spin hamiltonians. *arXiv preprint arXiv:1908.01745*, 2019.

[24] David Tritchler. An algorithm for exact logistic regression. *Journal of the American Statistical Association*, 79(387):709–711, 1984.

[25] Kenji Yano, Eiji Yamamoto, Koichiro Aya, Hideyuki Takeuchi, Pei-ching Lo, Li Hu, Masanori Yamasaki, Shinya Yoshida, Hidemi Kitano, Ko Hirano, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature genetics*, 48(8):927, 2016.

[26] 東芝デジタルソリューションズ 株式会社. 組合せ最適化ソルバー　simulated bifurcation machine. https://www.toshiba-sol.co.jp/pro/sbm/index_j.htm. 2021-01-11 参照.

[27] 岩井大介. 富士通とペプチドリーム、高速かつ高精度に中分子医薬候補化合物の探索を実現「デジタルアニーラ」活用で創薬プロセスを飛躍的に短縮、新型コロナウィルス感染症の治療薬開発への適用を目指す . https://pr.fujitsu.com/jp/news/2020/10/13c.pdf. 2021-01-24 参照.

[28] 久保拓弥. **データ解析のための統計モデリング入門一般化線形モデル・階層ベイズモデル・*MCMC***. 確率と情報の科学. 岩波書店, 2012.

[29] 大関真之 西森秀稔. **量子アニーリングの基礎**. 共立出版株式会社, 2018.

[30] 奥山　拓哉.　　最適化ソリューション　cmos アニーリングの活用事例. http://scsr.jp/document/20201124_SCSR_okuyama.pdf. 2021-01-11 参照.

[31] 日本電気株式会社.　Nec、量子コンピューティング領域に本格参入〜スーパーコンピュータを活用したアニーリングマシンによる共創サービスを提供開始〜. https://jpn.nec.com/press/201912/20191220_01.html. 2021-01-11 参照.

[32] 富士通株式会社. デジタルアニーラ. https://www.fujitsu.com/jp/digitalannealer/index.html. 2021-01-11 参照.