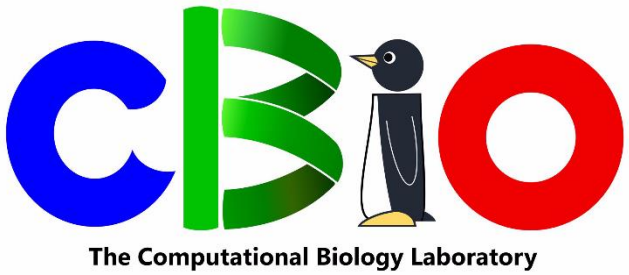
Computational Biology Laboratory



早稲田大学 情報理工学科
情報理工・情報通信専攻
清水研究室

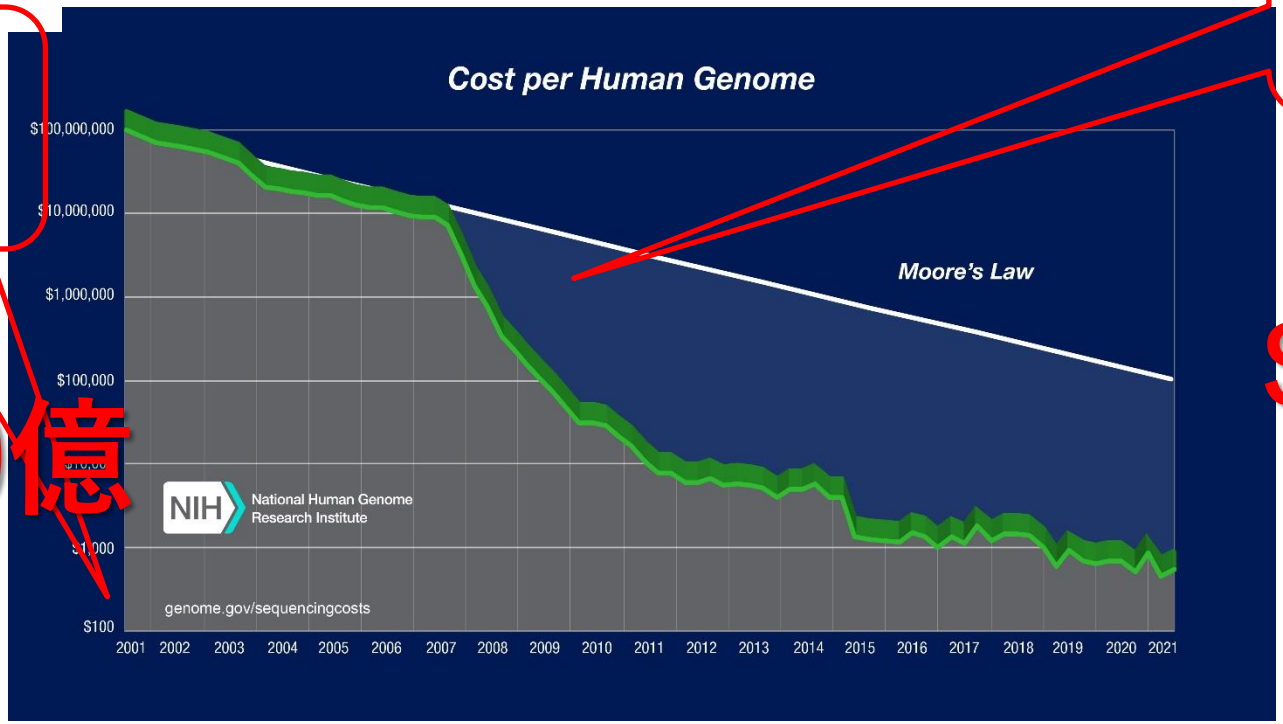
Sanger sequencer

https://products.appliedbiosystems.com

ゲノムビッグデータ時代の到来

ヒトゲノム計画
（1990～2003）

新型シークエンサーの実用化



$30億

$1000

http://jp.illumina.com/

Next Generation Sequencer

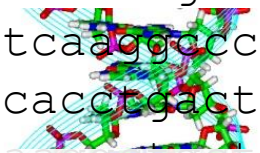
ヒトゲノムの長さ≒30億塩基対

生命の（ヒトであれば個人の）
設計図が容易に手に入るようになった！

ゲノムビッグデータ時代の生命科学

実験は大規模化，自動化
コンピュータによる解析が重要！

gtacaaaaaagcagaagg
gccgtcaaggcccaccat
ggtgcacctgactgatgc



生命の秘密を解き明かすデータ解析の手法の開発

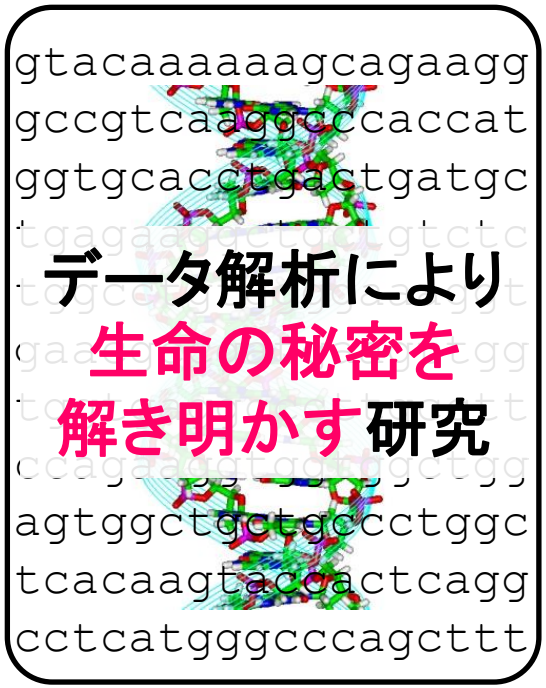
個人の設計図（ゲノム）を守るプライバシ保護技術の開発

$$\begin{aligned}
&\textbf{while } i < \ell \textbf{ do}\\
&\quad x = q[i+o]\\
&\quad p \leftarrow \mathrm{ExtR}(s,x)\\
&\quad \textbf{if } \|p\|_g > 0 \textbf{ then} \quad \triangleright \text{ Travers}\\
&\quad\quad b \leftarrow x
\end{aligned}$$

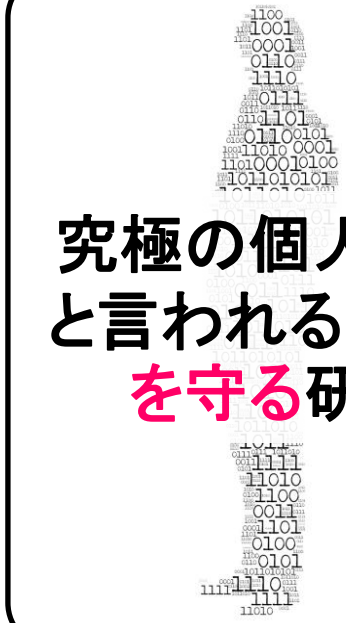
高度な分析を可能にする先端アルゴリズムの開発

研究室概要

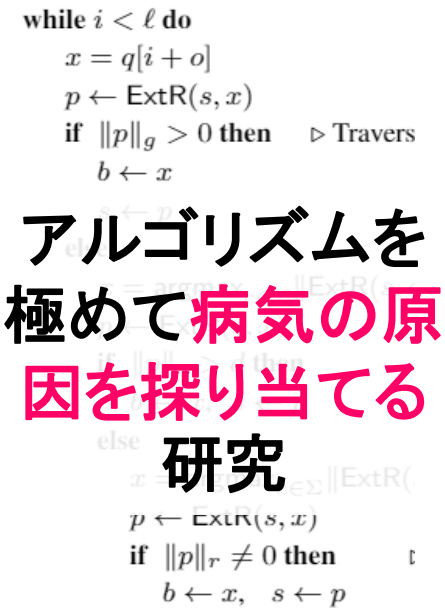
- 清水研では計算機科学のあらゆる手法を駆使して医学，生物学等の諸問題を解決する技術を研究します．



データ解析により**生命の秘密を解き明かす**研究



究極の個人情報と言われる**ゲノムを守る**研究



アルゴリズムを極めて**病気の原因を探り当てる**研究

Web: https://www.cbio.cs.waseda.ac.jp/

多くの研究テーマがありますが
そのうちの一部をご紹介！

生命情報科学×先端アルゴリズム

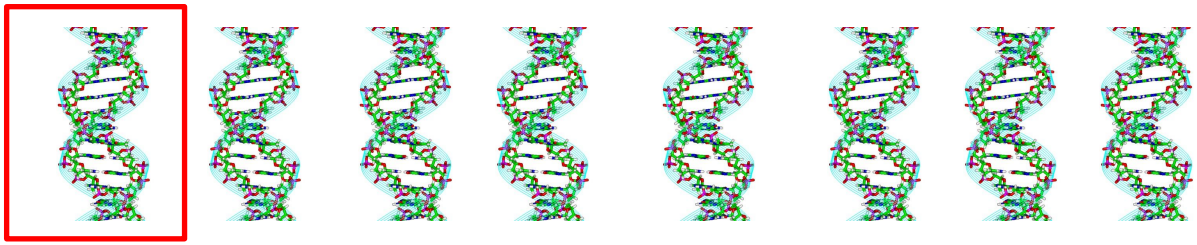
ゲノム配列解析では動的計画法や接尾辞配列等を用いた高度なアルゴリズムを多用します．高度過ぎるアルゴリズムは応用先がないと思われがちかもしれませんが，生命情報科学ではアルゴリズムの工夫が疾患等のメカニズム特定に役立つのです．

- **具体的にはどんな研究がある？**
  - 数万人規模のヒトゲノム配列を高速に検索可能な索引方法の研究

関連キーワード： 接尾辞木， 接尾辞配列， BWT, FM-Index, de Bruijn graph, 簡潔データ構造

参照ゲノムグラフ

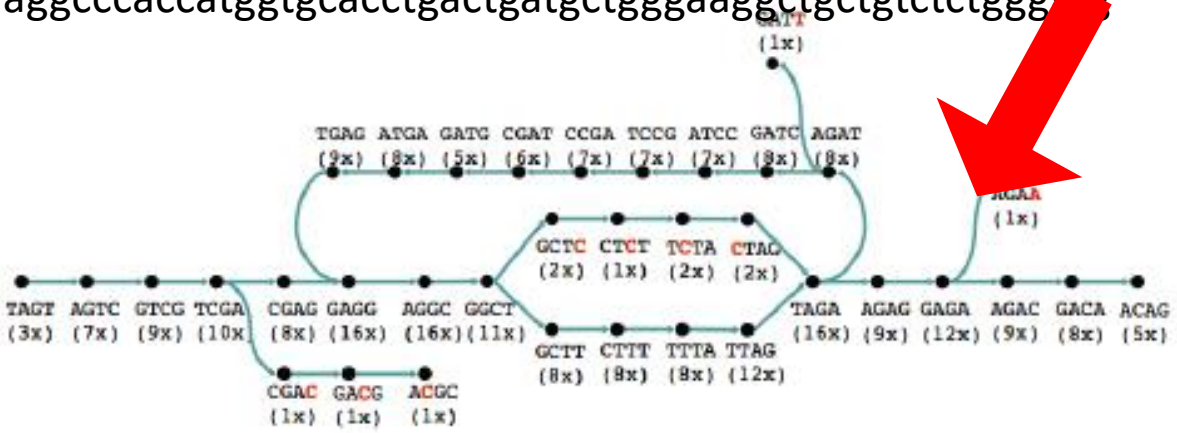
- 多様なゲノムを単一の辞書ではなく，グラフとして表現する新しい試み



一般化圧縮接尾辞配列, PBWT, GPBWTといった最先端のデータ構造の研究が必要！

gtacaaaaaagcagaagggccgtcaaggcccaccatggtgcacctgactgatgctgtgaaggctgctgtctctggcctg
gtacaaaaaagcagaaaaagccgtcaaggcccaccatggtgcacctgactgatgctgagaaggctgctgtctctggcct
gtacaaaaaagcagaagggccgtcaaggcccaccatggtgcacctgactgatgctgggaaggctgctgtctctgggc
....

多数の個人ゲノムを辞書化することによって，疾患等のメカニズムを正確に分析することができる．
しかし計算量等の技術的な課題がある！



http://plus.maths.org/content/os/issue55/features/sequencing/index, credit Daniel Zerbino

国際論文誌（JCB）に採録決定

色付きde Bruijnグラフに基づく索引のハッシュ関数を用いた精度向上とサイズ削減

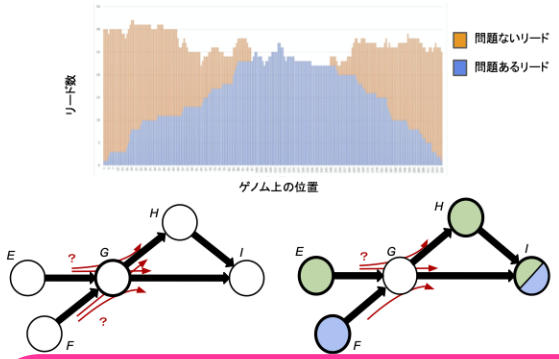
P-63 長谷川 望, 清水 佳奈 早稲田大学大学院基幹理工学研究科

1. 研究概要

- **リードの全長を全て保持したde Bruijnグラフ**を提案
  - **入力リード情報に即したノード遷移**が可能に
  - ゲノムアセンブリに基づく解析の精度を高められる
    - **がん原因遺伝子検出**の高精度化など
- 莫大なサイズのゲノムデータに対し，実用的なサイズ
  - Suffix Treeベースの索引と比べ，**サイズを改善**
- グラフへの付加情報を工夫し，**ノード遷移精度を改善**
  - **従来手法で特定の領域の情報が失われる問題を解決し，データの持つ情報をフル活用した解析が可能**に

2. 従来手法 [1]

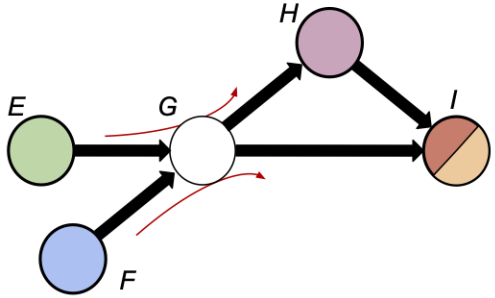
- de Bruijnグラフに色という情報を加え，入力配列のノード遷移を再現
- 1入力配列に対し，1色を割り当て
- **あるゲノム領域で正確なノード遷移を行えない問題**



3. 提案手法

色の割り当て方法の変更

- ハッシュ関数を用いて**配列の途中で色を変更**
- 現在のノードのIDと現在の色をキーとして，その次にノードに記録する色を求める
- 同じ配列由来のノードが同じ色をもつことを防ぎ，**従来手法の問題を解決**



色の記録方法の変更

- 色の記録に**ブルームフィルタ**[2] を用いる
- ノードごとに，飛んでくるクエリの数に応じてフィルタサイズを調整
- 各色を直接記録するよりも空間効率がよいため，**索引サイズの削減と省メモリ**につながる

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	1	0	0	0	0	0	1	1	0	0	1	0	0

偽陽性について

以下による偽陽性により，途中で辿れなくなる配列（**曖昧なリード**）が存在する可能性がある

- ハッシュ関数によって求めた色の値の衝突
- ブルームフィルタの誤判定

ただしいずれも**確率的に生じる**ので，従来手法よりもその後の**解析への影響は小さい**と考えられる

4. 実験結果

ヒト21番染色体のシミュレーションデータ
100塩基長×20,000,000配列を用いて索引構築

手法	索引サイズ [MB]	曖昧なリード数	実行時間 [second]
Suffix Tree	4,066	0	826
従来	487	122,938	3,624
提案1	2,653	0	2,829
提案2	2,155	629,159	3,194

5. 今後の展望

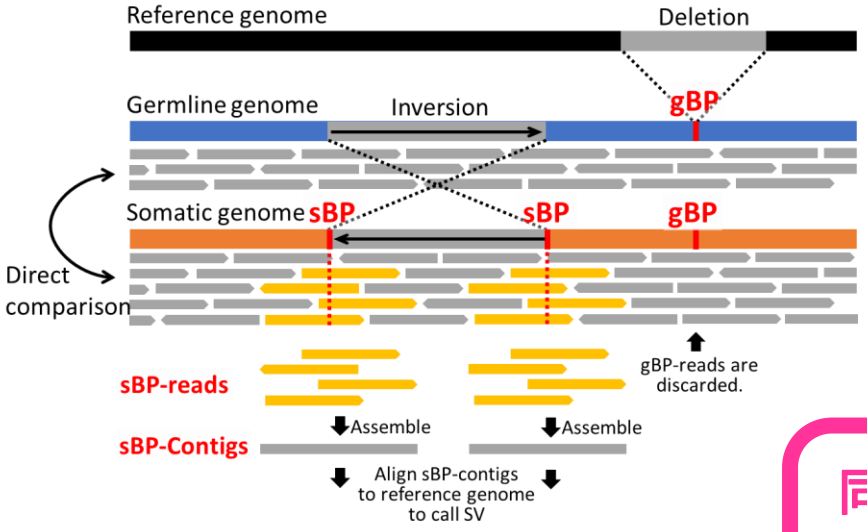
- より衝突性の低いハッシュ関数の実装や，色の記録方法の改良
- 全ゲノムシーケンスデータなど，よりサイズの大きなデータセットへの対応

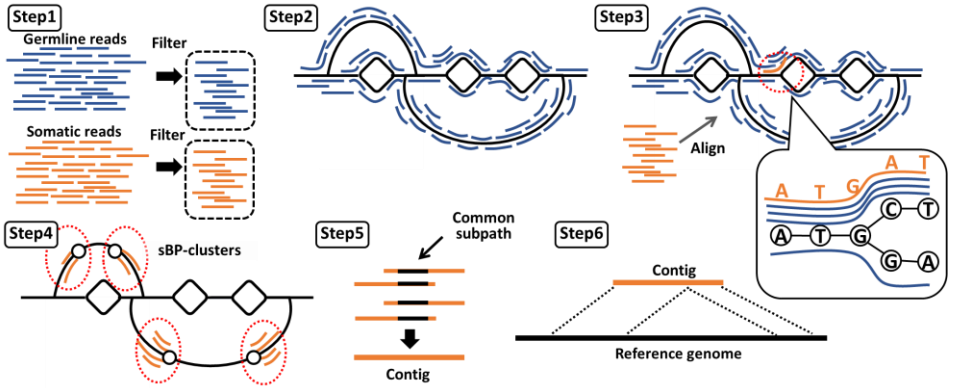
[1] D. Díaz-Domínguez, "An Index for Sequencing Reads Based on the Colored de Bruijn Graph.", International Symposium on String Processing and Information Retrieval, Springer, Cham (2019).
[2] Burton H. Bloom, "Space/time trade-offs in hash coding with allowable errors." Communications of the ACM 13.7, (1970).

メモリサイズを増加させることなく，精度よく情報を保存．従来，分析不可能であったゲノム領域の分析が可能に！

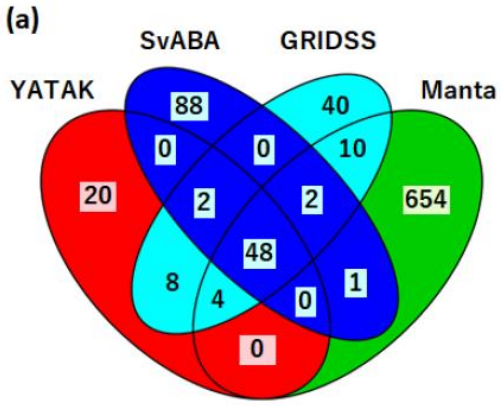
※ IIBMP 優秀ポスター賞受賞

ガンゲノム配列解析ソフトウェアの開発





同一人物のがん細胞と健康な細胞のゲノムを精度よく比較し，従来法では発見できなかった変異を発見可能に！



生命情報×セキュリティ

究極の個人情報を守るにはどうしたらよいでしょうか？ゲノム情報解析の知識とセキュリティの知識を総動員して難題に挑みます.

- **具体的にどんな研究ができる？**
  - 秘密計算技術を使ってゲノム情報を守りながら分析
- **どんな力が身につく？**
  - セキュリティ関連の技術を勉強できる他，ゲノム配列をはじめとした生命データの扱い方について知識を深めることができます.

関連キーワード：個人ゲノム，個人情報保護，プライバシ保護データマイニング，暗号プロトコル，準同型暗号，秘密分散，ORAM，Trusted execution environment

プライバシ保護データマイニング（Privacy-preserving datamining）

- 個別の**データの中身を見ないまま解析を行い必要な情報のみを抽出する技術**の総称 [Agrawal+2000, Lindell+2000]
- 例: 教員の給与総額を計算

  （通常の計算の場合）個別の給与額を見て計算

プライバシ保護データマイニング（Privacy-preserving datamining）

- 個別の**データの中身を見ないまま解析を行い必要な情報のみを抽出する技術**の総称 [Agrawal+2000, Lindell+2000]
- 例: 教員の給与総額を計算

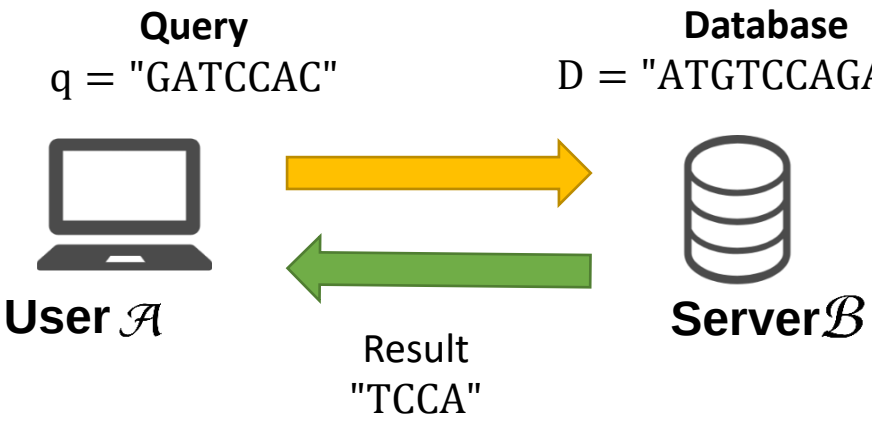
（通常の計算の場合）個別の給与額を見て計算

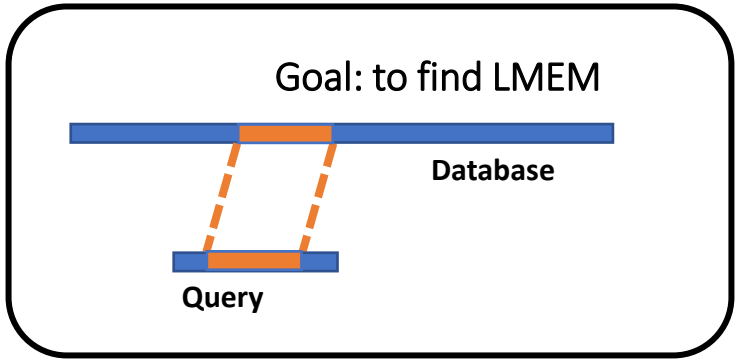
（PPDMの場合）個別の給与額を見ないで計算

秘密全文検索法 (CSS2019, WABI2021)

- 短い文字列（クエリ，約100文字程度）と長い文字列（DB，約数万～）間の**最長一致部分文字列**を発見する．ゲノム配列検索など，応用範囲は広い．
- 事前計算で用意したシェア[1]の利用により，**クエリ投入から検索結果を得るまでの時間／空間／通信量が定数オーダーのテーブル参照法**を開発．（モデルはclient aided 2P）
- O(L)で全文検索が可能な手
- パターンマッチ，木構造検

データを秘密にしたまま，ゲノム配列の比較ができるように．データの大きさに依存しない超効率的な方法を開発！





1) Correlated randomnessでクエリに依存せずに使える．

CSS 研究奨励賞受賞

学習器の安全な利用 [Sudo+2018]

決定木の出力を秘匿計算する効率的手法の提案.

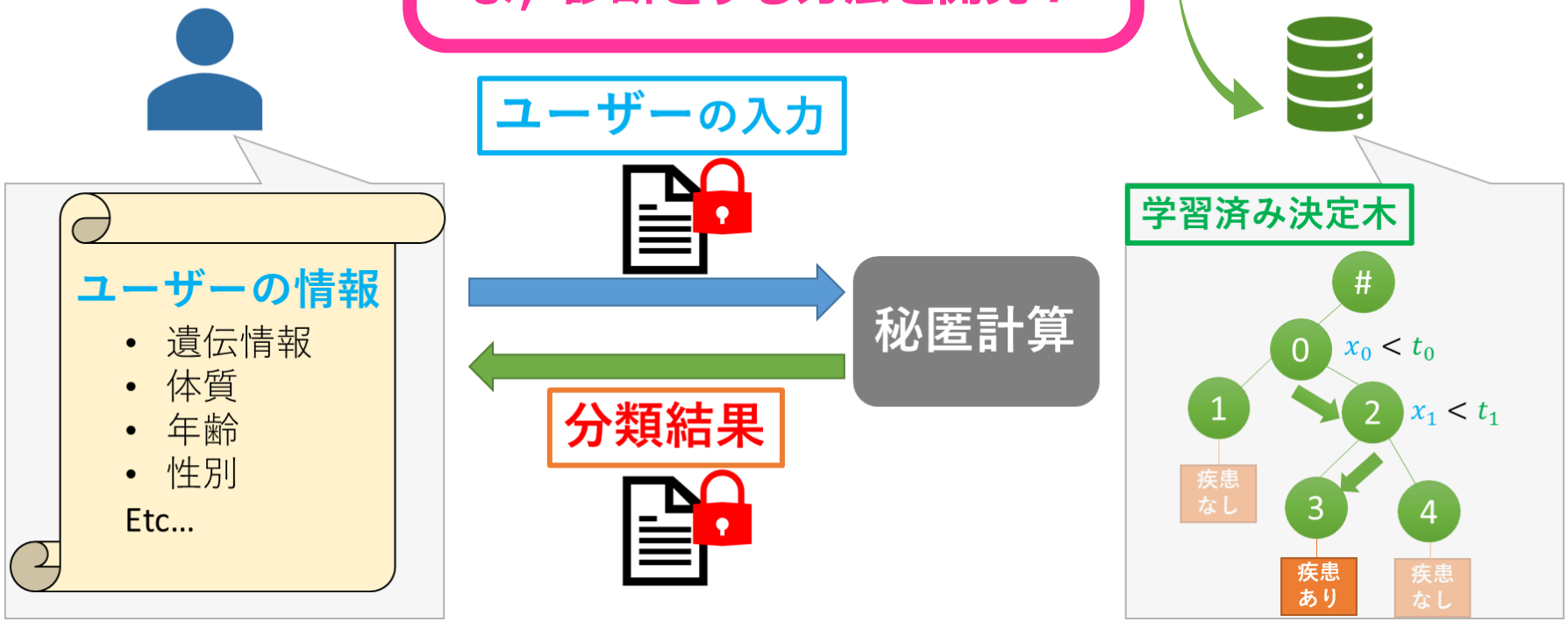
- ユーザーは決定木の出力のみを得る
- サーバーは何

ユーザー

ユーザーの個人情報と機械学習のモデルを互いに秘密にしたまま，診断をする方法を開発！

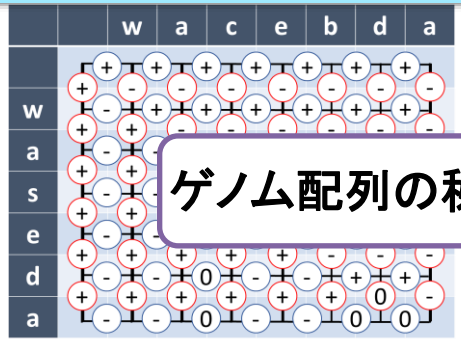
サーバー

学習



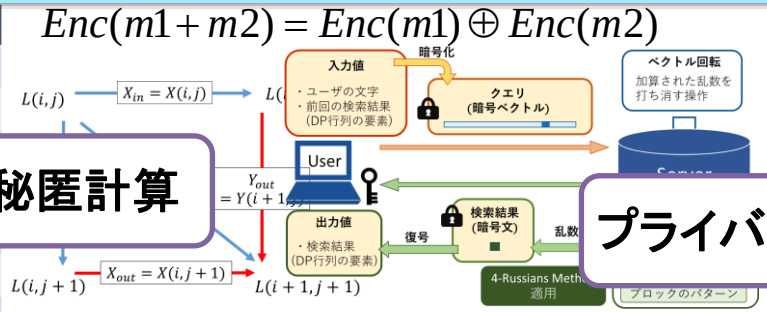
DEIM学生プレゼンテーション賞

卒論テーマ例

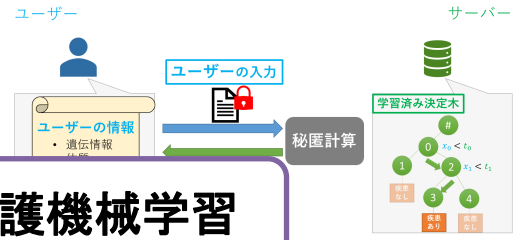


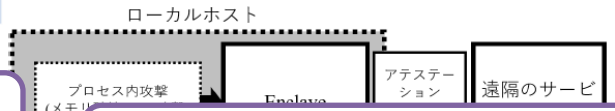
ゲノム配列の秘匿計算

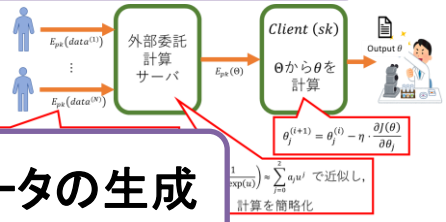
$$Enc(m1+m2) = Enc(m1) \oplus Enc(m2)$$



プライバシ保護機械学習

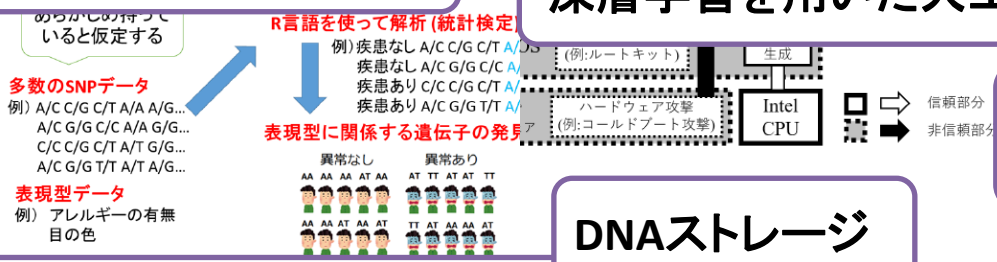




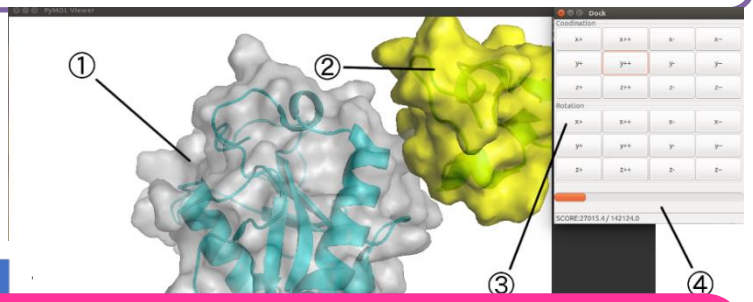


ゲノムワイド関連解析

深層学習を用いた人工ゲノムデータの生成

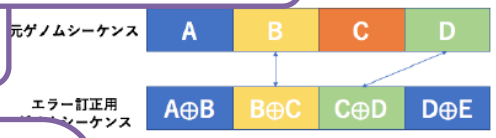


ゲーミフィケーションによる生体分子ドッキングシミュレーション



DNAストレージ

効率的なゲノムグラフの構築

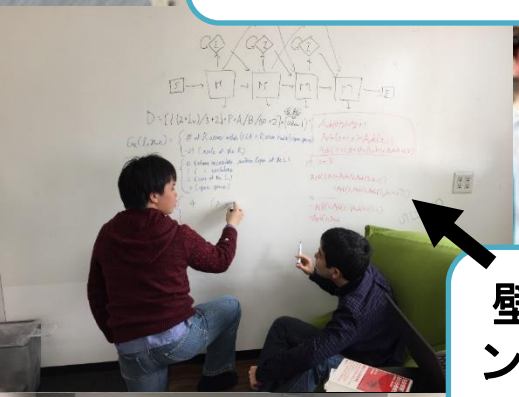


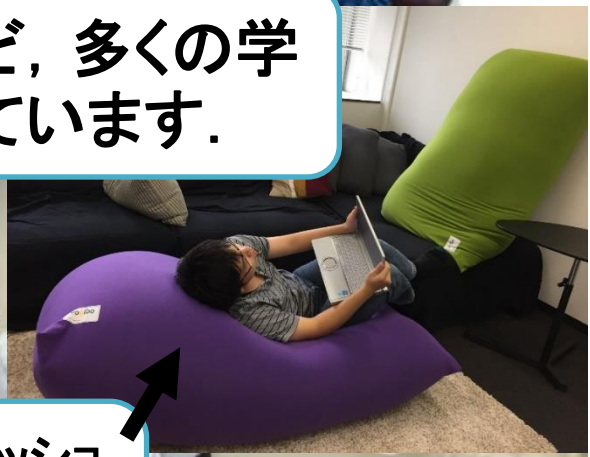
DTCゲノミクスの安全性
Spaced Seed検索のデータ構造アルゴリズム
など

研究テーマは自由に選択／提案できる
色々な課題に挑戦

研究室の様子

国内外での研究発表や学会での受賞など，多くの学生が大学外からも認められる活躍をしています．





壁一面のホワイトボードや大きなビーズクッションがあります．キッチンのある部屋では，勉強会が開かれ，活発な議論が繰り広げられます．

研究室で重視していること

- **よく考え，議論をして思考力を養う**
  - 論理的な思考力は，一生モノの財産.
  - 教員や共同研究先のプロの研究者との対話を通して確かな実力を養う
  - 勉強会，日々のゼミでの議論
- **チャレンジ精神を養うこと**
  - 新しい研究テーマへのチャレンジ，各自のやりたいことを応援します.
- **学内にとどまらずに活躍すること**
  - 国内外の学会発表，国際ハッカソン参加
- 自分で考え，実践し，その結果を世に出すことの**楽しさを実感してほしい**.（そのサポートを全力でします）

研究室の参考動画・記事

【研究室紹介動画・パンフレット】 ※ 研究室の様子，学生インタビュー等．

- パンフレット：http://www.waseda.jp/nyusi/ebro/ug/se_jp_2019/html5.html#page=15
- 動画：https://www.youtube.com/watch?v=VmsU9H3eGPU

【研究内容紹介動画・記事】

- IPSJONE(https://ipsj-one.org/2016/)講演:
  https://ipsj-one.org/2016/videos/24_shimizu_fs.mp4
- 河合塾・みらいぶプラス掲載記事:
  生命情報科学について: https://www.milive-plus.net/gakumon161202/02/
  ゲノム情報保護について: https://www.milive-plus.net/gakumon161202/
- 学部web（生命情報科学について）:
  http://www.fse.sci.waseda.ac.jp/june-1-2016/
- WASEDA ONLINE（ゲノムビックデータ解析について）:
  https://yab.yomiuri.co.jp/adv/wol/opinion/science/20230619.php
- 早稲田理工 by AERA（秘匿ゲノム検索について）:
  https://www.cbio.cs.waseda.ac.jp/assets/materials/AERA_KS.pdf
- CAMPUS NOW 10月号（秘匿ゲノム検索について）:
  https://waseda.box.com/s/h3n2ogcy1k8ivxz5oqnbpnm8s07dakga
- JSPS記事（ゲノムビックデータ解析について）:
  https://www.jsps.go.jp/j-grantsinaid/22_letter/data/news_2018_vol1/p12.pdf

The Computational Biology Laboratory

その他

- オープンハウスは，対面，オンラインの双方で実施します．
  詳しくは，web（https://www.cbio.cs.waseda.ac.jp/docs/posts/）をご参照ください．
- 研究環境等

  全員に対して個人用の席とPCが割り当てられます．テーマや研究の進捗にもよりますが，学外の一流研究者との議論も活発に行います．
- 学生の活躍

  国内外での研究発表や学会での受賞など，多くの学生が大学外からも認められる活躍をしています．
- 研究室の所在，連絡先

  63号館5階01, 21室

  E-mail：shimizu.kana@waseda.jp

  URL: https://www.cbio.cs.waseda.ac.jp