TOKYO TECH
Pursuing Excellence

Tokyo Institute of Technology –TOKYO TECH– develops distinctive students with outstanding qualities of creativity and leadership. TOKYO TECH is making significant contributions to science and technology in many fields of expertise, creating new and powerful synergies. TOKYO TECH, being a research-based university, is dedicated to education and research, and to exploring knowledge in science and technology. Pursuing excellence, TOKYO TECH serves society and the world.

SCLS
SUPERCOMPUTATIONAL LIFE SCIENCE

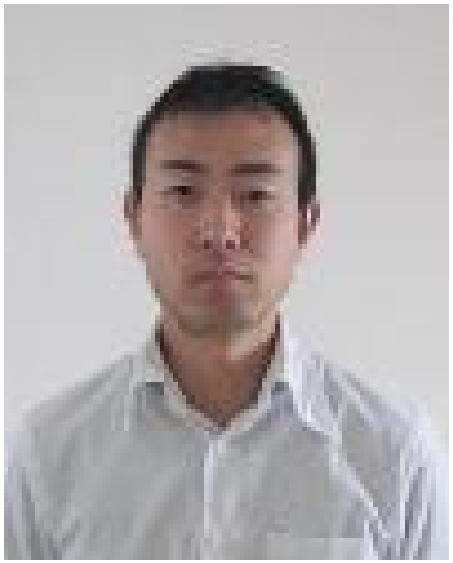
並列配列相同性検索プログラム「GHOST-MP」講習会（講義編）

2015年3月20日

東京工業大学　大学院情報理工学研究科
角田 将典、石田 貴士、秋山 泰

講師紹介

角田　将典
かくた　まさのり



石田　貴士
いしだ　たかし

秋山　泰
あきやま　ゆたか

東京工業大学　大学院情報理工学研究科　計算工学専攻

本日の予定

- 13:00-13:05　ごあいさつ
- 13:05-13:50　GHOST-MP講習
- 13:50-14:00　休憩
- 14:00-16:00　GHOST-MP実習

関連文献紹介

- GHOST-MP関連文献
  - GHOSTX: Suzuki et al., (2014) *PLoS ONE* 9(8):e103833
    - 接尾辞配列を用いたアラインメント候補位置の高速探索
  - GHOST-MP: Kakuta et al., (in preparation)
    - GHOSTXの分散メモリ環境版、
- 当グループの他の配列相同性検索関連文献
  - GHOXTM: Suzuki et al., (2012) *PLoS ONE* 7(5): e36060
    - GPUを用いた相同配列検索
  - GHOSTZ: Suzuki et al., (in press) doi: 10.1093/bioinformatics/btu780
    - 部分文字列のクラスタリングによるアラインメント候補位置の高速探索
  - GHOSTZ-GPU: Suzuki et al., (in preparation)
    - GHOSTZのGPU版

アジェンダ

- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

GHOST-MPとは

- 配列相同性検索プログラム
  - 塩基配列やアミノ酸配列をクエリ、アミノ酸配列を検索対象とする
  - 感度が高く、高速な検索
    - GHOSTXアルゴリズム（Suzuki et al. 2014）による高速な検索
    - Message Passing Interface （MPI）とOpenMPによる並列化による計算資源の利用
    - 大量クエリ配列の並列検索を高速に行える
      - 1本のクエリ配列からなる検索では、恩恵は小さい

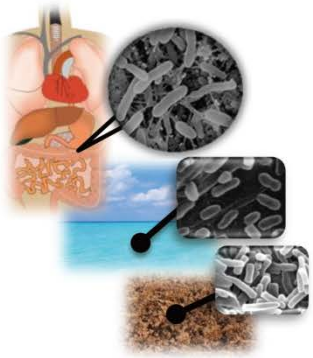
アジェンダ

- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

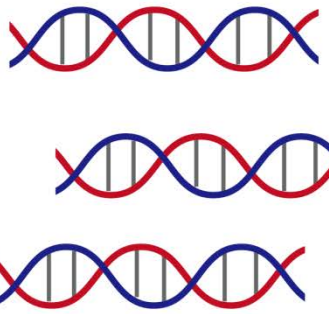
環境と細菌叢

- ヒトをはじめとして動物の体表・体内や、土壌、海洋などの環境中には様々な微生物が存在する
- 同じ環境内でも微生物集団（細菌叢）には多様性があり、環境と細菌叢は相互に影響を与えている
  - ヒト腸内の細菌叢同士を比べても、条件（個人、疾病、乳児の成長過程など）によって、細菌の組成が異なる
- 環境と細菌叢の関係を調査するため、環境中の細菌叢の情報を明らかにする必要がある

環境中の細菌叢のDNA Sequencingによる解析（1）



サンプルの取得



DNA の抽出

...CCTTATCTTCG...

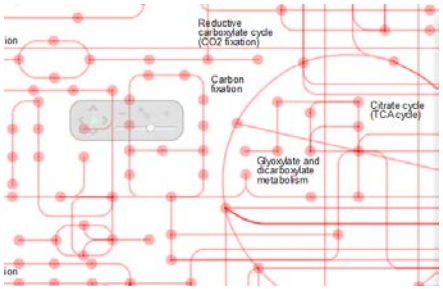
...CCACATAAACT...

...ATGGTCGATGTT...

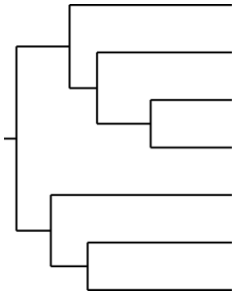
塩基配列の読み取り

塩基配列から様々な解析が可能

分類群・遺伝子の
相対存在度による解析



パスウェイ解析



系統樹解析

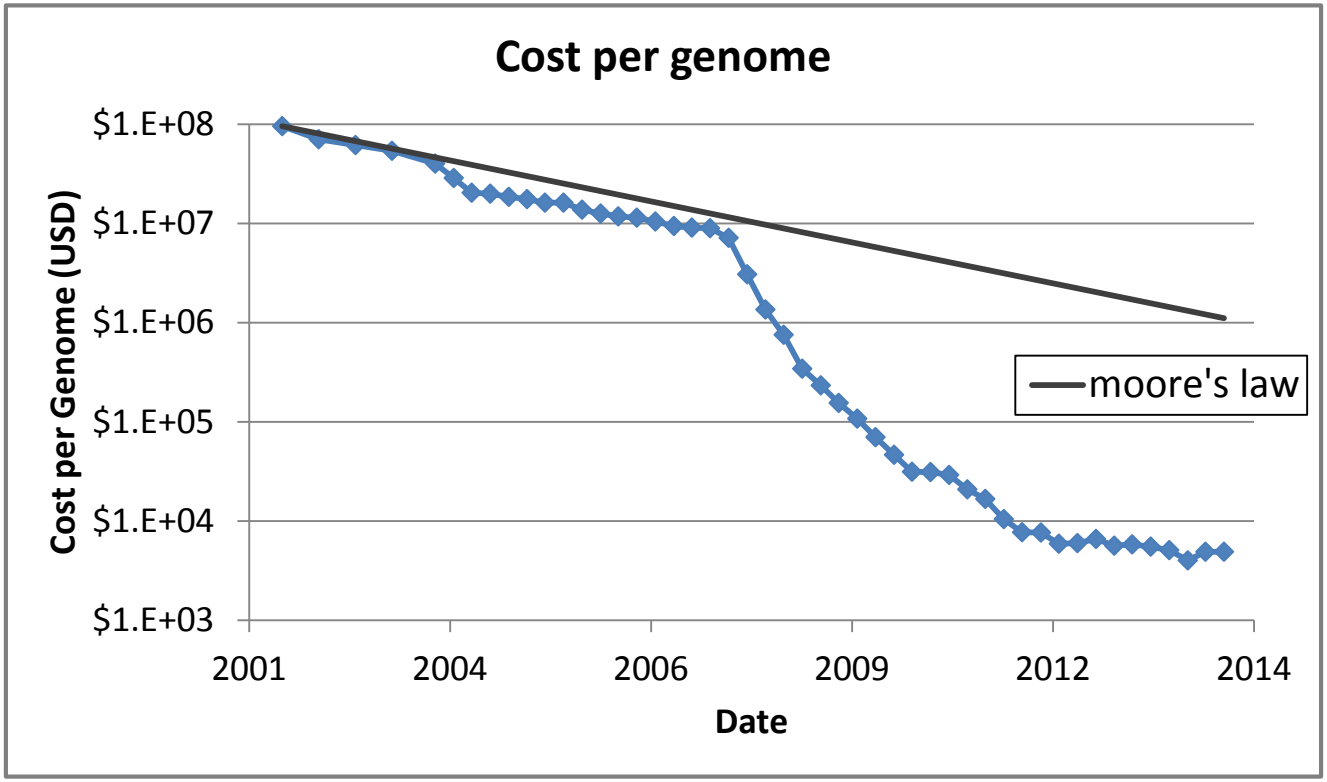
環境中の細菌叢のDNA Sequencingによる解析（2）

- **マーカー遺伝子（16S rRNAなど）**
  - 特定の遺伝子がsequencingの対象
    - 対象がマーカー遺伝子に限られるため、必要なシーケンシングデータは小さい
  - どのような細菌がどのくらい存在するか解析
- **メタゲノム**
  - 細菌叢の全ゲノムがsequencingの対象
    - 全ゲノムが対象であるため、必要とされるシーケンシングデータが大きい
  - どのような細菌がどのくらい存在するか解析
  - どのような遺伝子がどのくらい存在するか解析

- シーケンサの性能向上によって可能になった
- メタゲノムデータの解析では、配列解析の対象となる配列数と塩基数が大きいため、高速な解析が要求される

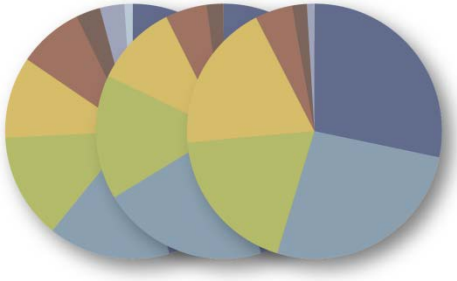
DNA Sequencingの近年の傾向

DNA Sequencingコストの推移（ヒトゲノム）

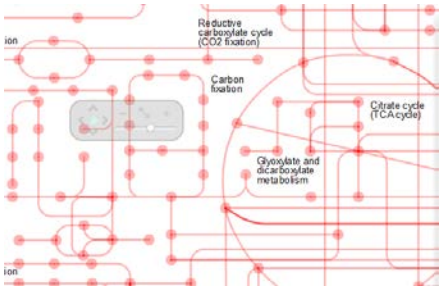


Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed Jan 10, 2015.

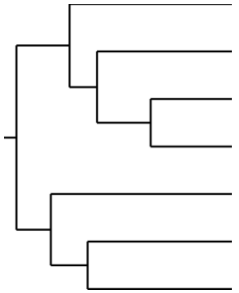
配列相同性検索が解析で果たす役割



分類群・遺伝子の
相対存在度による解析



パスウェイ解析



系統樹解析

- 配列相同性検索は、読み取った塩基配列の由来する分類群や遺伝子ファミリ、機能などの推定に用いられる
  - 塩基配列のみでは、分類群や遺伝子に関する情報は不明
  - 配列相同性検索により、既知の類似配列を探し、それらを推定する

GHOST-MPの開発動機

- メタゲノム解析の際の配列相同性検索に、多くの時間を要する

クエリ: 土壌メタゲノムのシーケンシングデータ (75bp x 72M reads)
NGS system (Illumina GAII)
DB: NCBI nr (about 5GB)
KEGG genes.pep (about 2GB)

NCBI BLASTX
on 144-core Intel Xeon PC cluster

約400時間

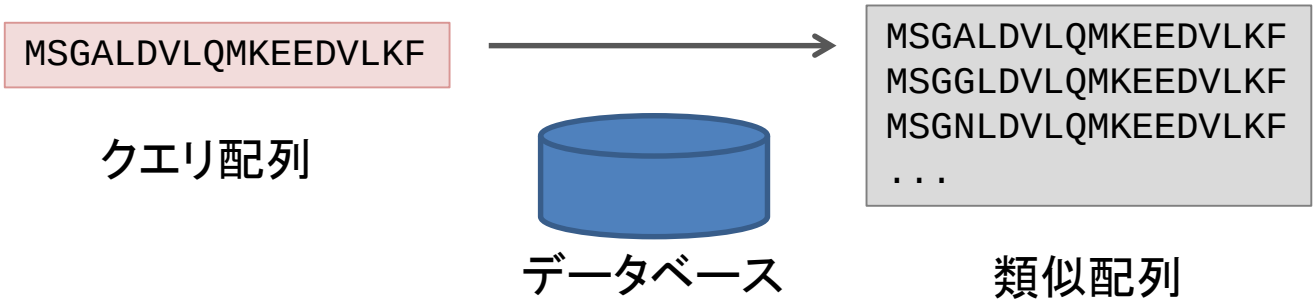
高速な配列相同性検索が必要とされる

アジェンダ

- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

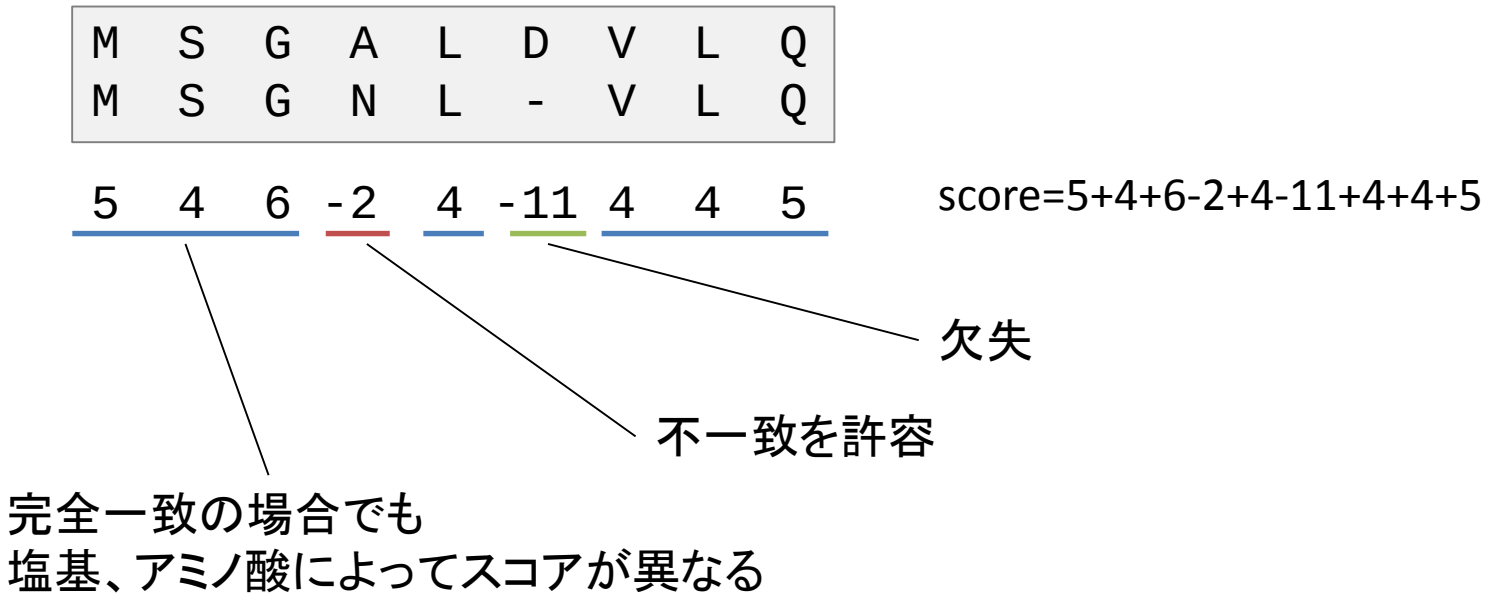
配列相同性検索

- 進化的に類縁関係にある配列（相同配列）、つまり、共通の祖先を有する配列では、機能が保存していると推定することができる
- 配列相同性検索は、相同配列としてデータベースから類似配列を検索する手法



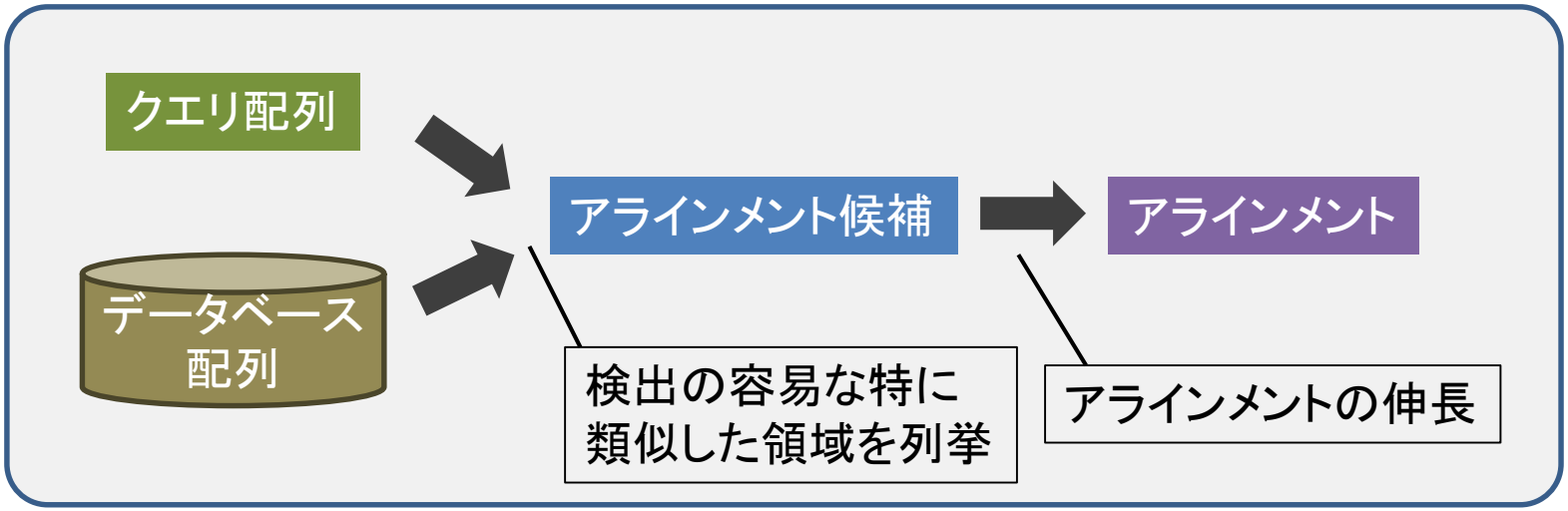
配列相同性検索（配列の類似性）

- **塩基またはアミノ酸の類似性、挿入、欠失を考慮してアラインメントし、スコアを評価する**

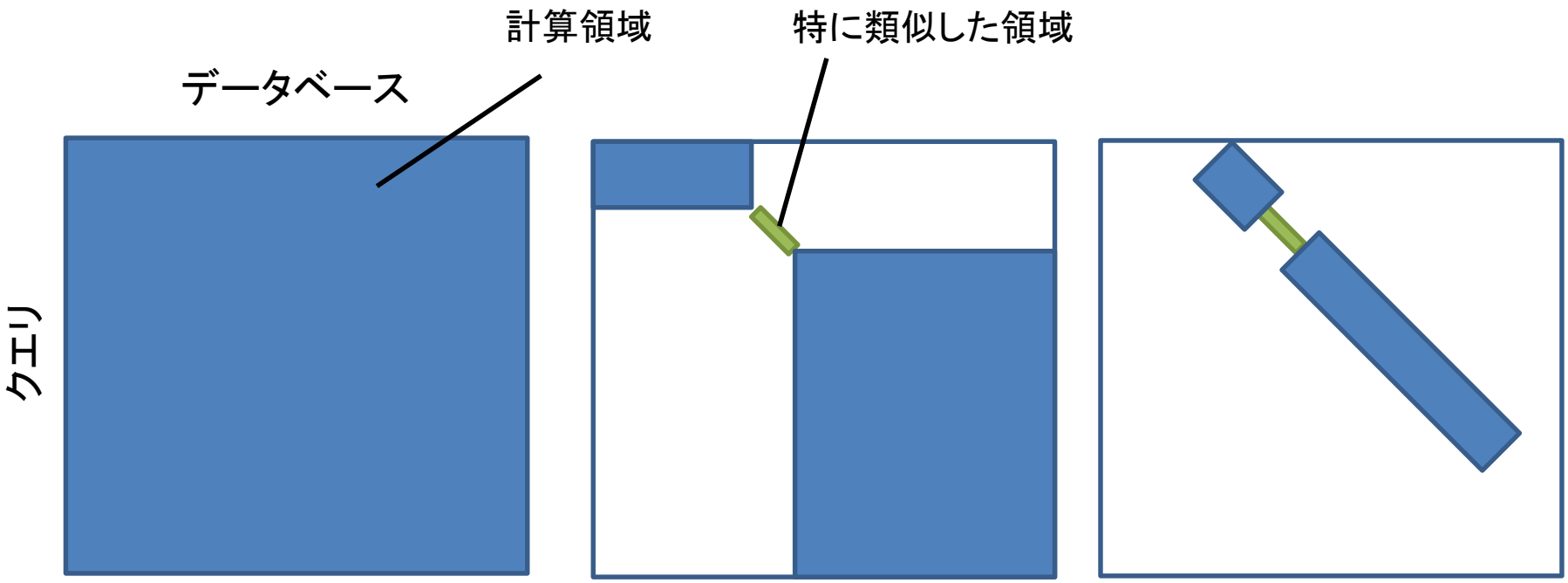


配列相同性検索(候補探索)

- 様々な方法が提案されている基本的には、類似配列の検索時間を短縮するため、高速に候補を探索した後、候補についてアラインメントの評価を行う



配列相同性検索(候補探索)



Smith-Watermanなどで
最適解を求める場合

特に類似した領域を見つけ、
その部分のアラインメントを
確定することで計算領域を
削減できる

類似スコアが低くなった際に
挿入・欠失の伸長を打ち切る
ことで、計算領域をさらに
削減できる

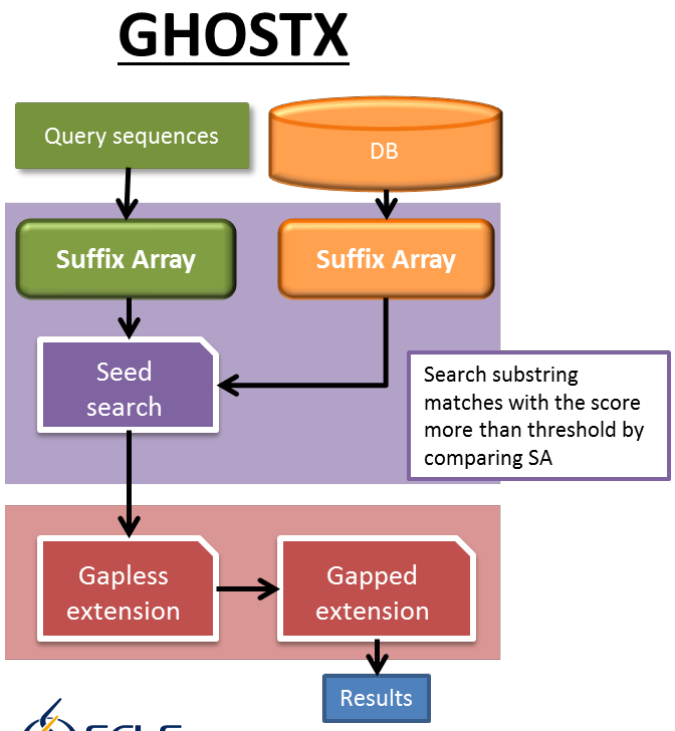
アジェンダ

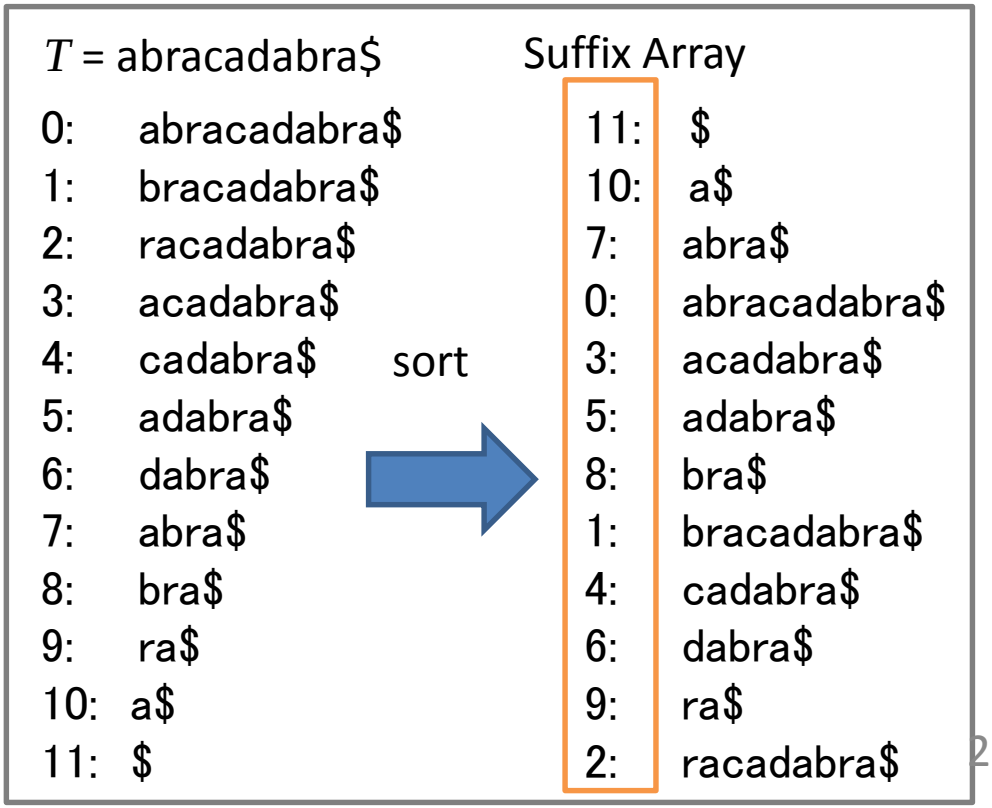
- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

GHOSTXアルゴリズム（1）

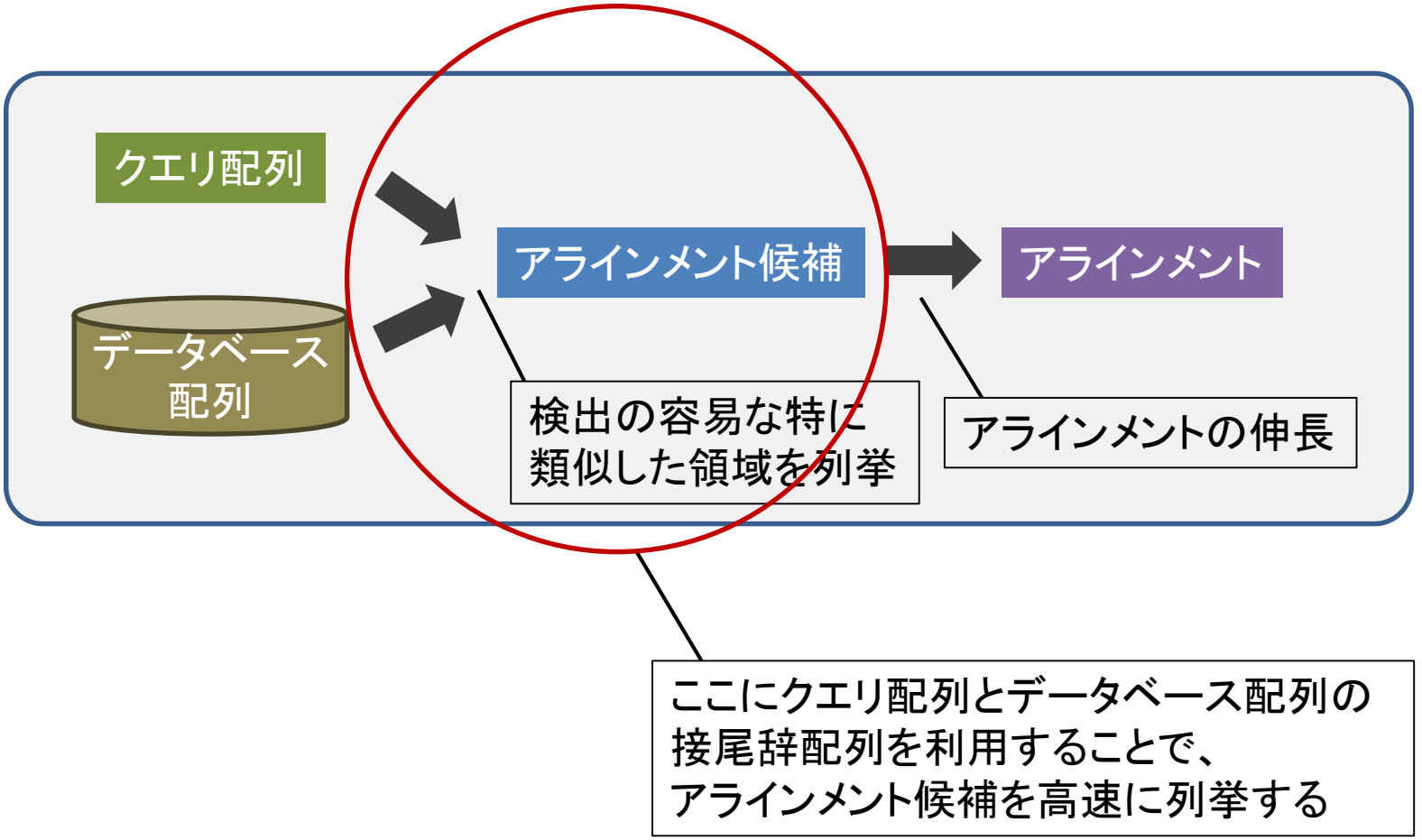
Suzuki et al. (2014) *PLoS ONE* 9(8):e103833

- アラインメント候補位置を高速に探索するアルゴリズムを提案し、
  これによって高速な相同性検索を実現した
  - 接尾辞配列（Suffix Array）というデータ構造を用いて、
    二分探索を行うことでクエリとデータベースの一部を比較するだけで、
    候補位置を見つけることができる。配列全てを突き合わせて比較しないため高速

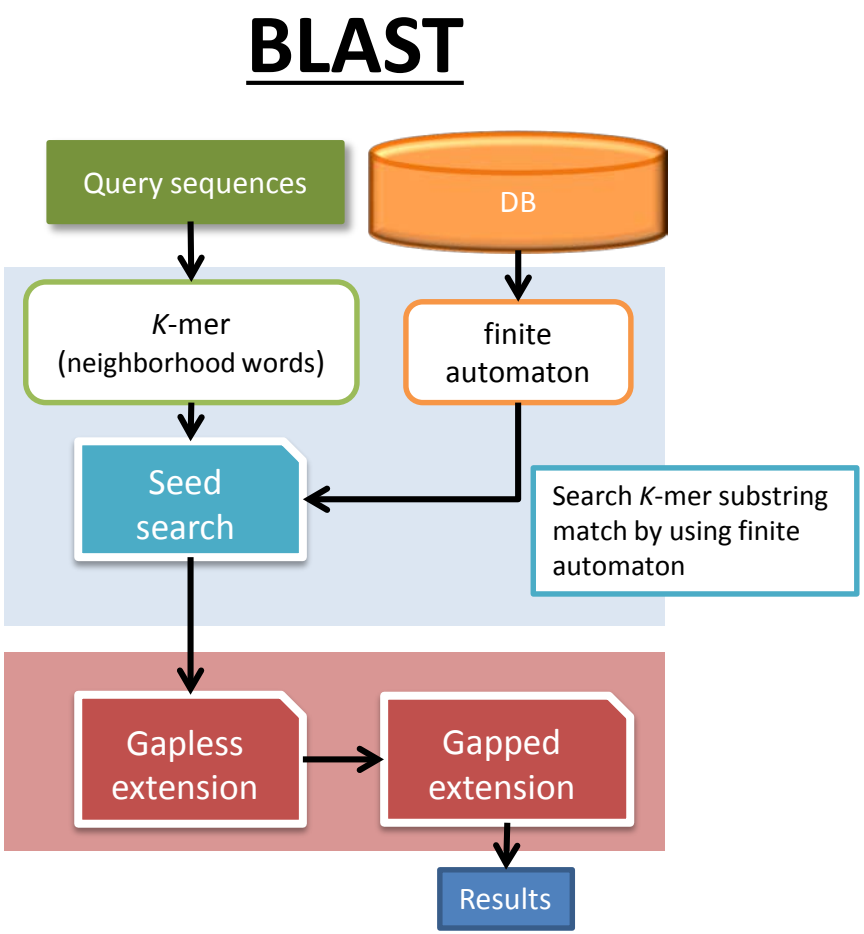


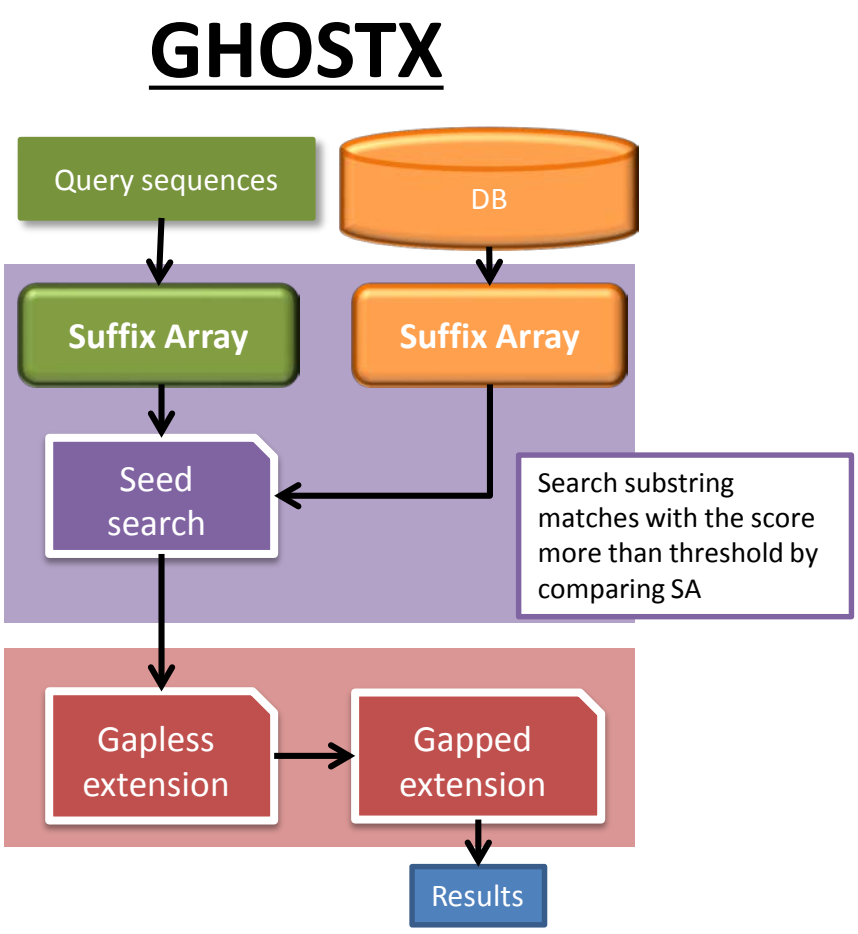


GHOSTXアルゴリズム（2）



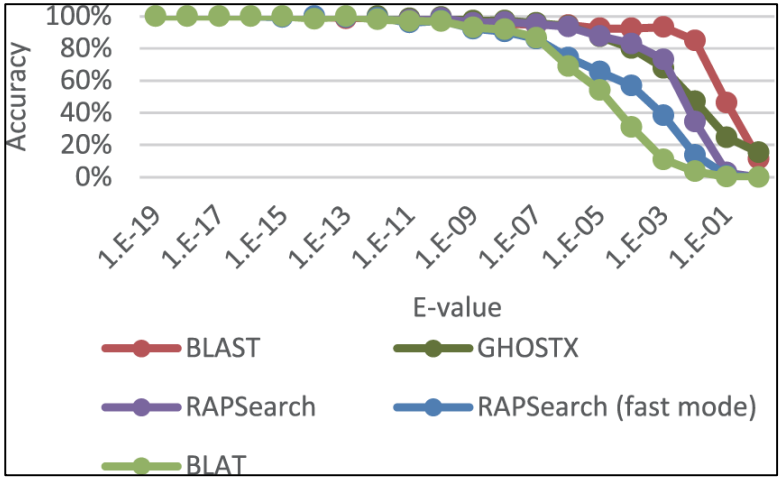
GHOSTXアルゴリズム（3）





GHOSTXの精度と速度

- 計算ノード1ノード、1スレッドを利用した場合
- BLASTと比較し152倍高速
- 近年開発されメタゲノム解析に用いられているRAPSearchと比較しても、同等の精度で高速に検索が行えた



	Computation time (sec.)	Acceleration ratio
GHOSTX	401.9	152.6
RAPSearch	649.5	94.4
RAPSearch in fast mode	91.2	672.2
BLAT	1409.7	43.5
BLAST	61314.1	1.0

The first, second, and third columns show the name of each program, the computation time, and the acceleration in processing speed relative to BLASTX using 1 thread, respectively.
doi:10.1371/journal.pone.0103833.t001

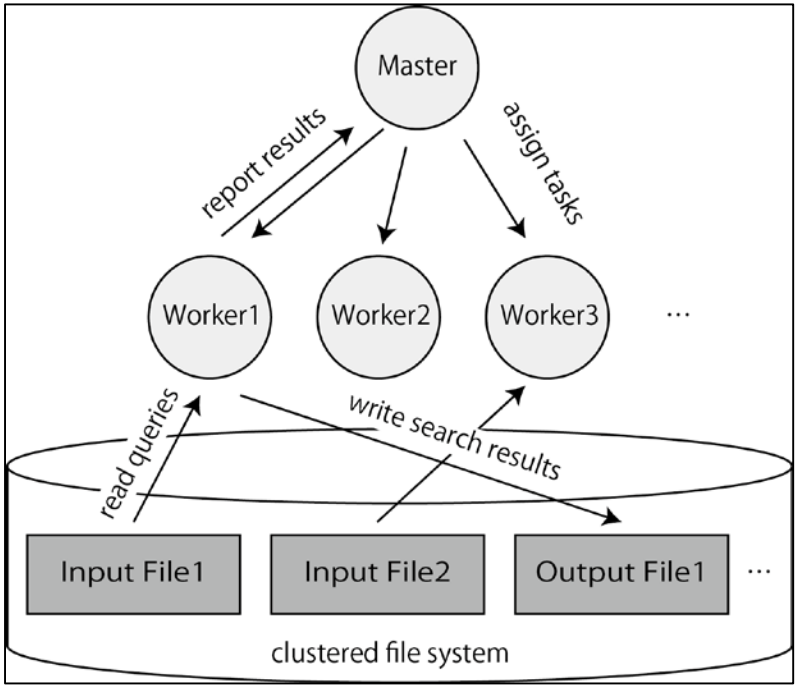
アジェンダ

- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

GHOST-MP

(Kakuta et al. in preparation)

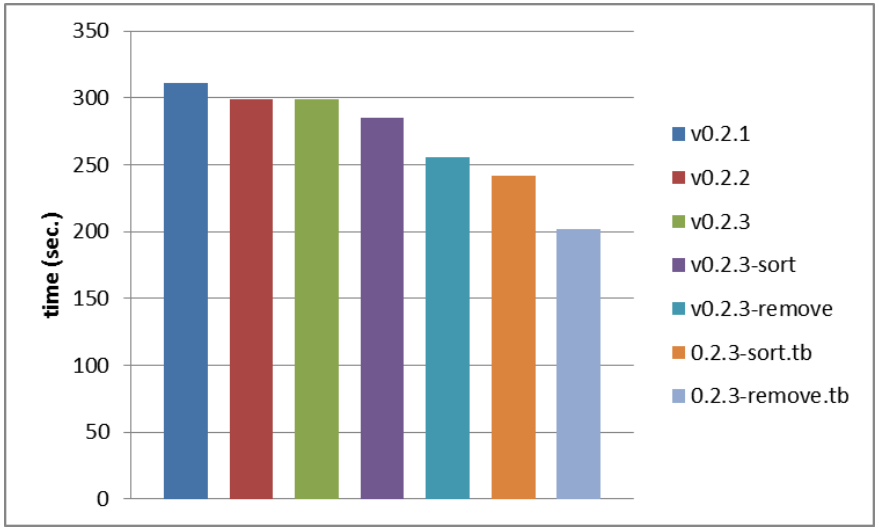
- GHOSTXアルゴリズムを用いて複数の計算ノード上で大規模並列検索を行う
  - 特にスパコン「京」で実行することを念頭に開発
- スパコンをはじめとして近年の計算機の高速化は計算ユニット（コア、ソケット、ノード）の増加によって行われているため並列計算に対応することは重要
- 分散メモリ環境では計算ノード間でデータが共有できないため、ノード間のデータ移動をMPIを実装した



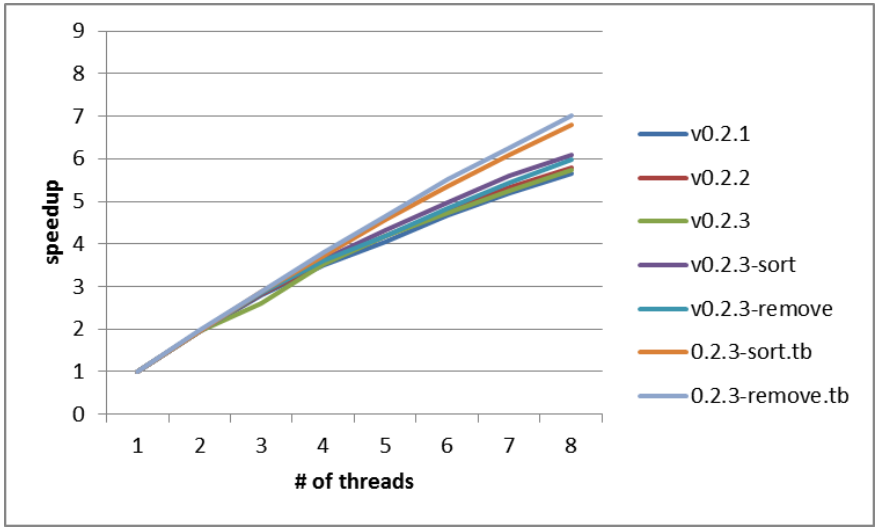
GHOST-MP

- GHOSTXアルゴリズムの「京」の計算環境に対する最適化
  - メモリの確保・メモリアクセスの最適化
  - スレッド間の負荷分散の改善

プログラム全体



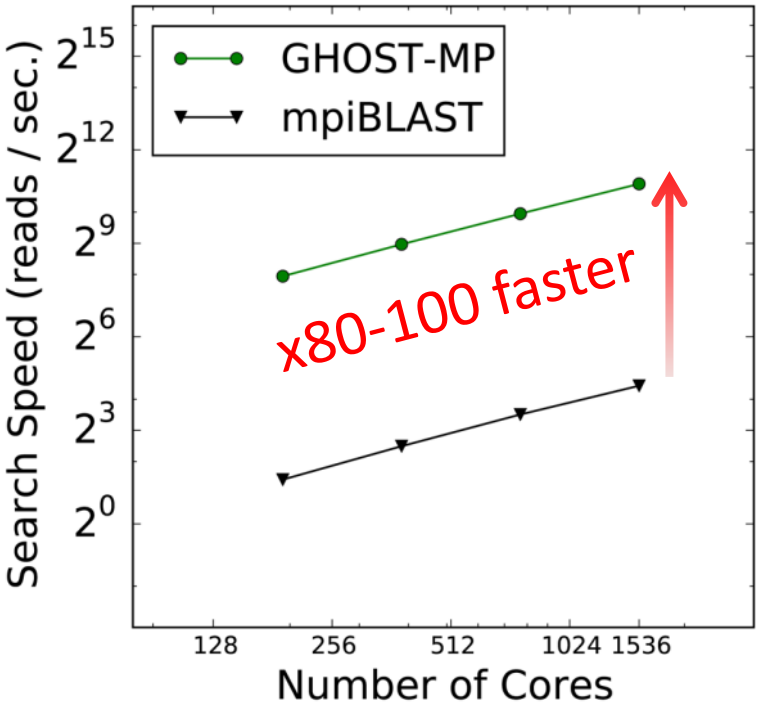
経過時間



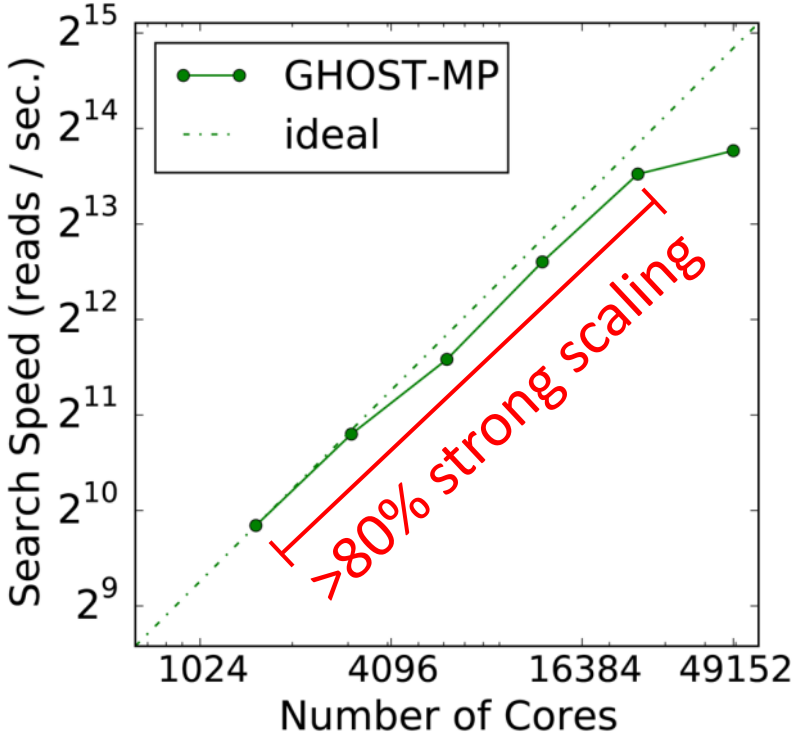
1スレッド使用時に対する速度向上

GHOST-MP

- 検索アルゴリズム自体はGHOSTXと同じため、精度に変化はない
- BLASTの並列実装であるmpiBLASTと比較し、同じ計算機資源を用いて80-100倍高速であった
- 「京」を用いた実験で使用コアの増加と共に32,000 CPUコアまで計算速度が向上



Strong scaling on TSUBAME 2.5

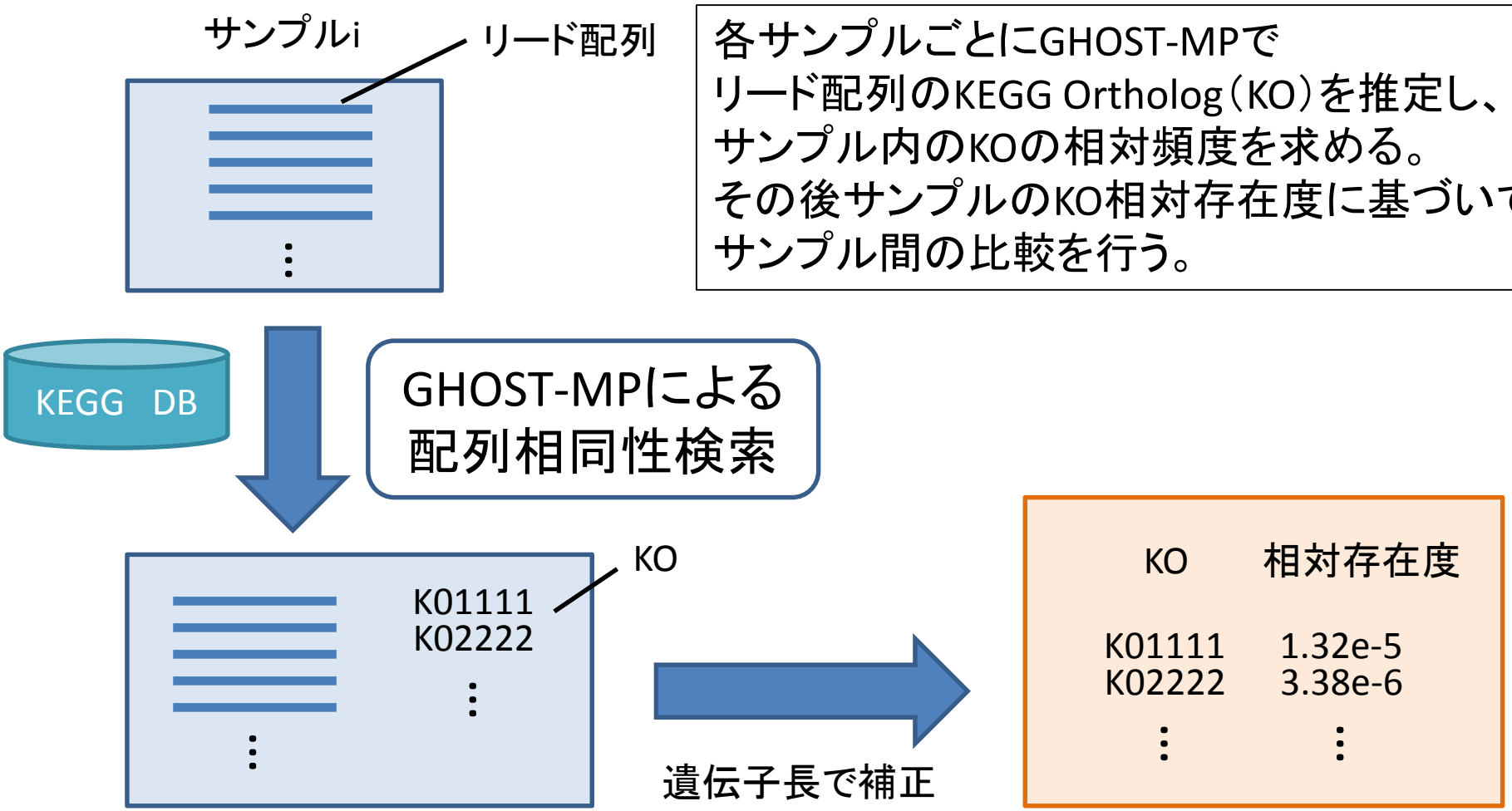


Strong scaling on K computer

アジェンダ

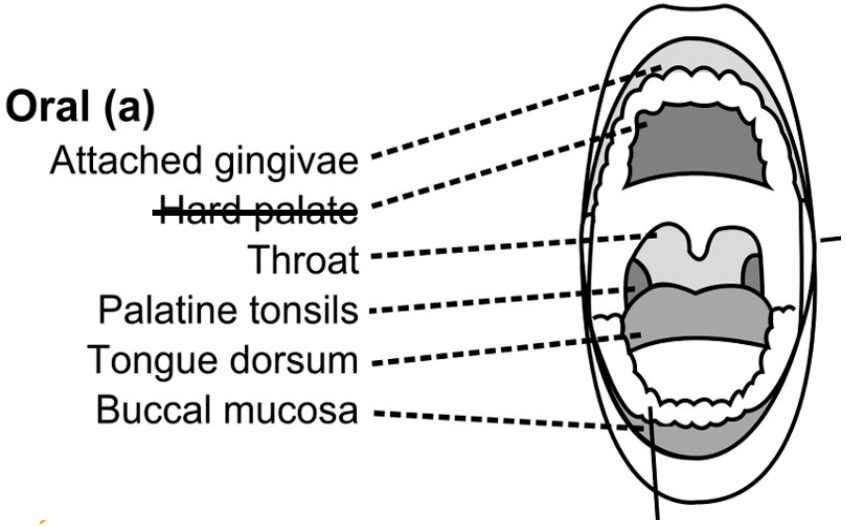
- GHOST-MPとは
- GHOST-MPの開発動機
  - メタゲノム解析
- 配列相同性検索
- GHOSTXアルゴリズム
- MPIによる分散メモリ環境での並列化
- メタゲノム解析（GHOST-MPの応用として）

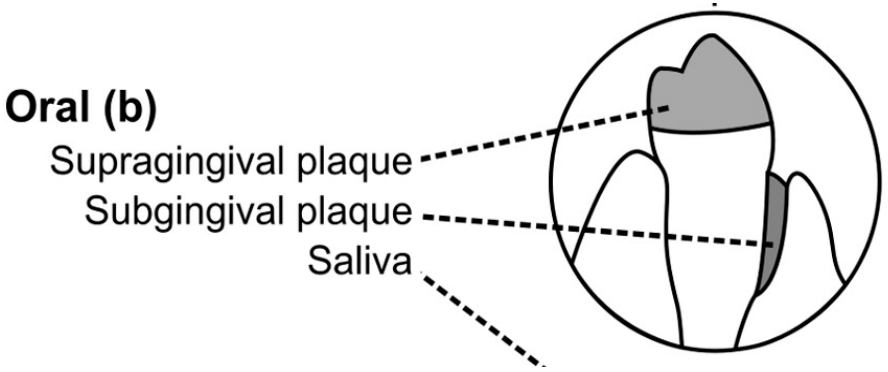
解析処理の概要


サンプルi
リード配列
各サンプルごとにGHOST-MPで
リード配列のKEGG Ortholog（KO）を推定し、
サンプル内のKOの相対頻度を求める。
その後サンプルのKO相対存在度に基づいて
サンプル間の比較を行う。
KEGG DB
GHOST-MPによる
配列相同性検索
KO
K01111
K02222
遺伝子長で補正
KO
相対存在度
K01111
1.32e-5
K02222
3.38e-6

ヒト口腔内メタゲノム解析への応用

- GHOST-MPを利用してHuman Metagenome Project（HMP）の公開するシーケンシングデータの解析を行った
  - 口腔内8部位、381サンプル、180億リード
  - 部位： 角化歯肉、~~硬口蓋~~、咽喉、口蓋扁桃、舌背、頬粘膜、歯肉縁上の歯垢、歯肉縁下の歯垢、唾液





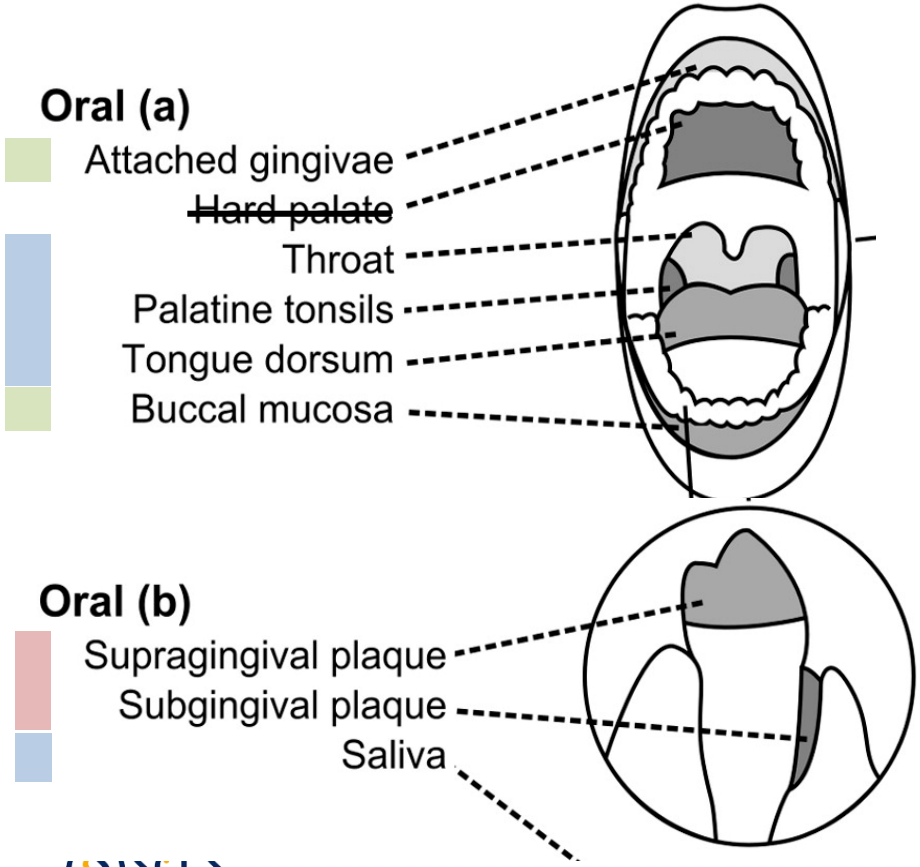
L.M. Proctor, Cell Host & Microbe, 2011

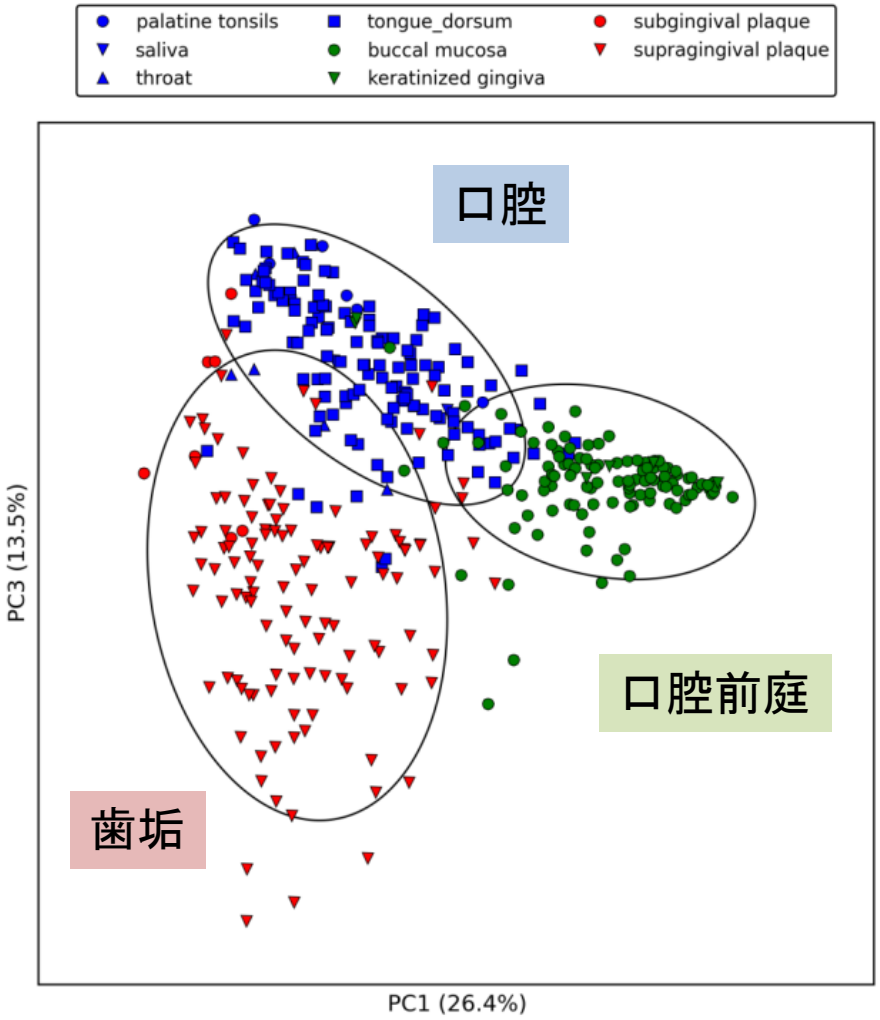
HMP口腔メタゲノムデータ内訳

Site	# of samples	# of reads (x 10^6)
角化歯肉	6	331
硬口蓋	0	0
咽喉	7	128
口蓋扁桃	6	129
舌背	127	10330
頬粘膜	107	1202
歯肉縁上の歯垢	118	6200
歯肉縁下の歯垢	7	137
唾液	3	23
Total	381	18484

KO相対存在度によるサンプル間比較

- 主成分分析を行った
- 第3主成分までで、58%の累積寄与率
- 第1、第3主成分で口腔内、口腔前庭、歯垢のデータの分布が異なることが分った





L.M. Proctor, Cell Host & Microbe, 2011