並列配列相同性検索プログラム「GHOST-MP」講習会（実習編）

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

実習の目的

- GHOST-MPの利用方法の習得
  - GHOST-MPの基本的な利用方法
  - SCLS計算機システム上での利用
    - スケジューラを介した実行方法
- GHOST-MPによるメタゲノムデータに対する相同性検索結果の解析
  - 分類群に基づく解析
  - 遺伝子オーソログに基づく解析

実習前の確認

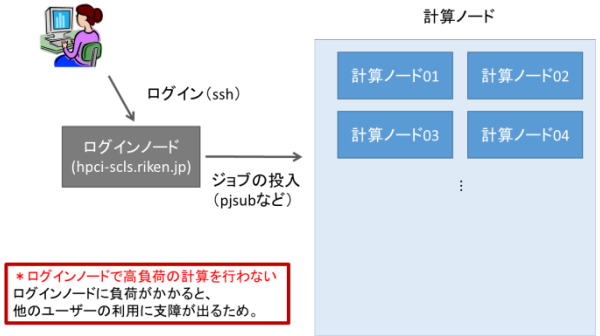
- SCLS計算機システムにログインできること
- ログインノードにおける基本的なファイル操作ができること
- ウェブブラウザでAdobe Flash Playerが利用可能なこと
  - ＊GHOST-MPの実行とは直接関係はないが、GHOST-MPの検索結果の可視化の際に必要

アジェンダ

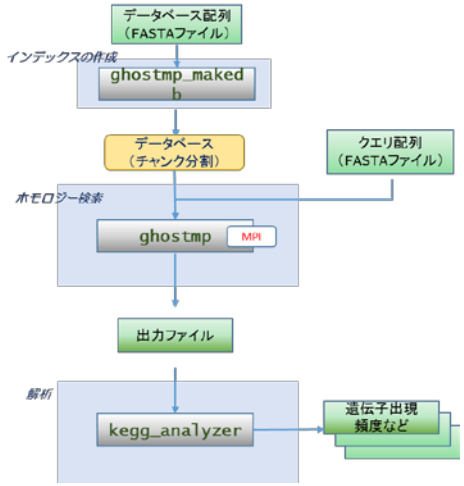
- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

本日の実習内容

SCLS計算機システム上でのGHOST-MPの利用

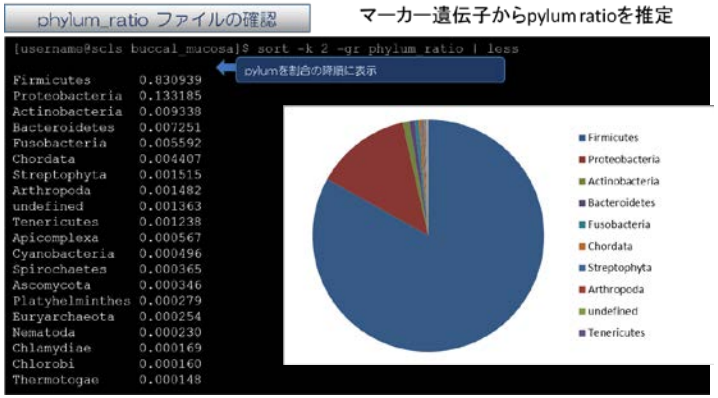


SCLS計算機システムの利用方法

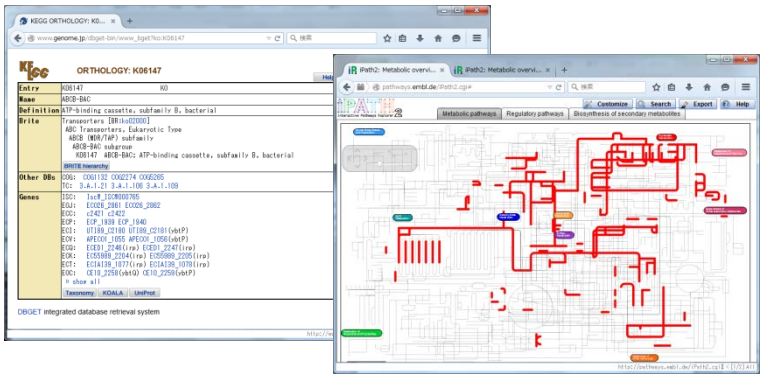


GHOST-MPの利用方法

GHOST-MPのメタゲノムデータに対する相同性検索結果の解析



分類群に基づく解析



遺伝子・機能に基づく解析

GHOST-MP

BLAST のように遠縁のホモログを検出可能な配列相同性検索ツールである
GHOSTXアルゴリズムを並列化したもの

GHOST-MP

- MPIとOpenMPによるハイブリッド並列化
  MPIによる計算ノード間の並列化、OpenMPによる計算ノード内の並列化を行い、クエリ配列のデータ並列処理を行う
- ファイル I/O
  MPI-IOを利用し、並列入出力を行う
- 事前にデータベースのインデックス作成が必要
  GHOSTXアルゴリズムは、高速な検索を実現するために、接尾辞配列（Suffix Array）というデータ構造を事前に作成しておく

関連文献

- GHOST-MP関連
  - 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
  - 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

GHOST-MP（ハイブリッド並列化）

検索処理全体

MPIによる並列化

OpenMPによる並列化

- アラインメント候補の探索
  - 接尾辞配列による類似部分の探索
  - アラインメントスコアの評価

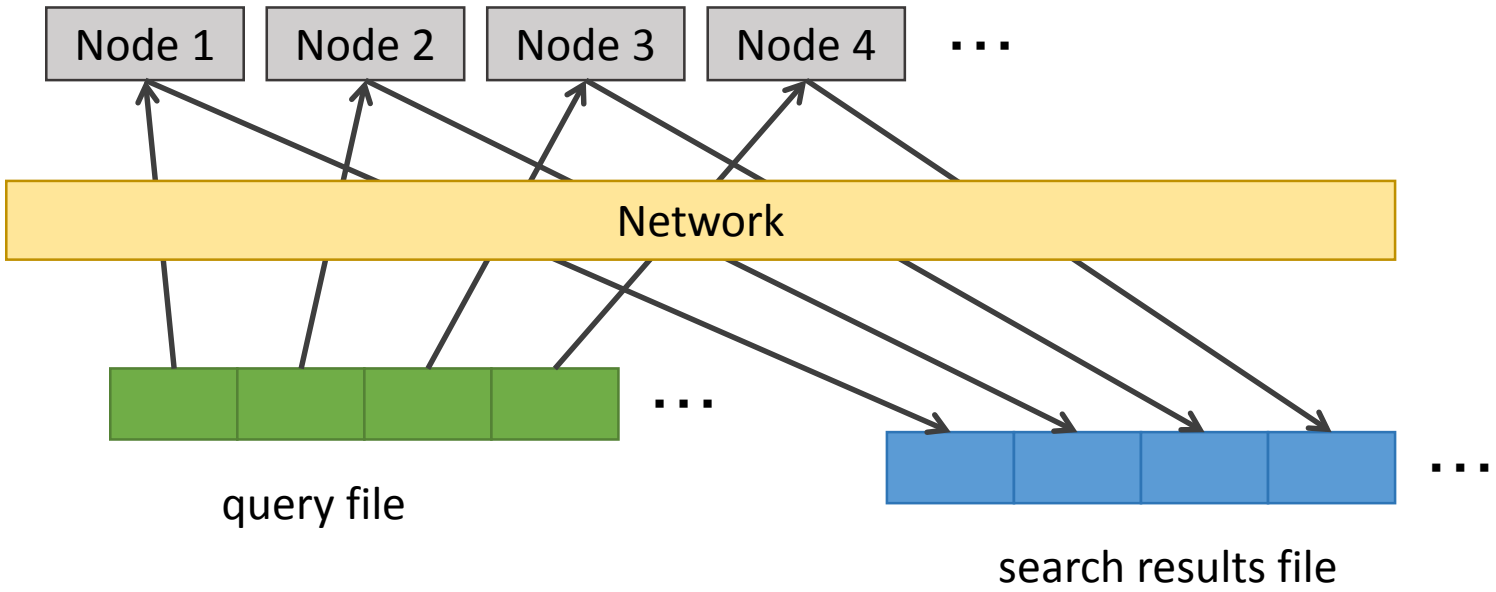
OpenMPによる並列化

- アラインメント作成

- ファイルI/O

GHOST-MP

MPI-IOによる並列ファイル入出力

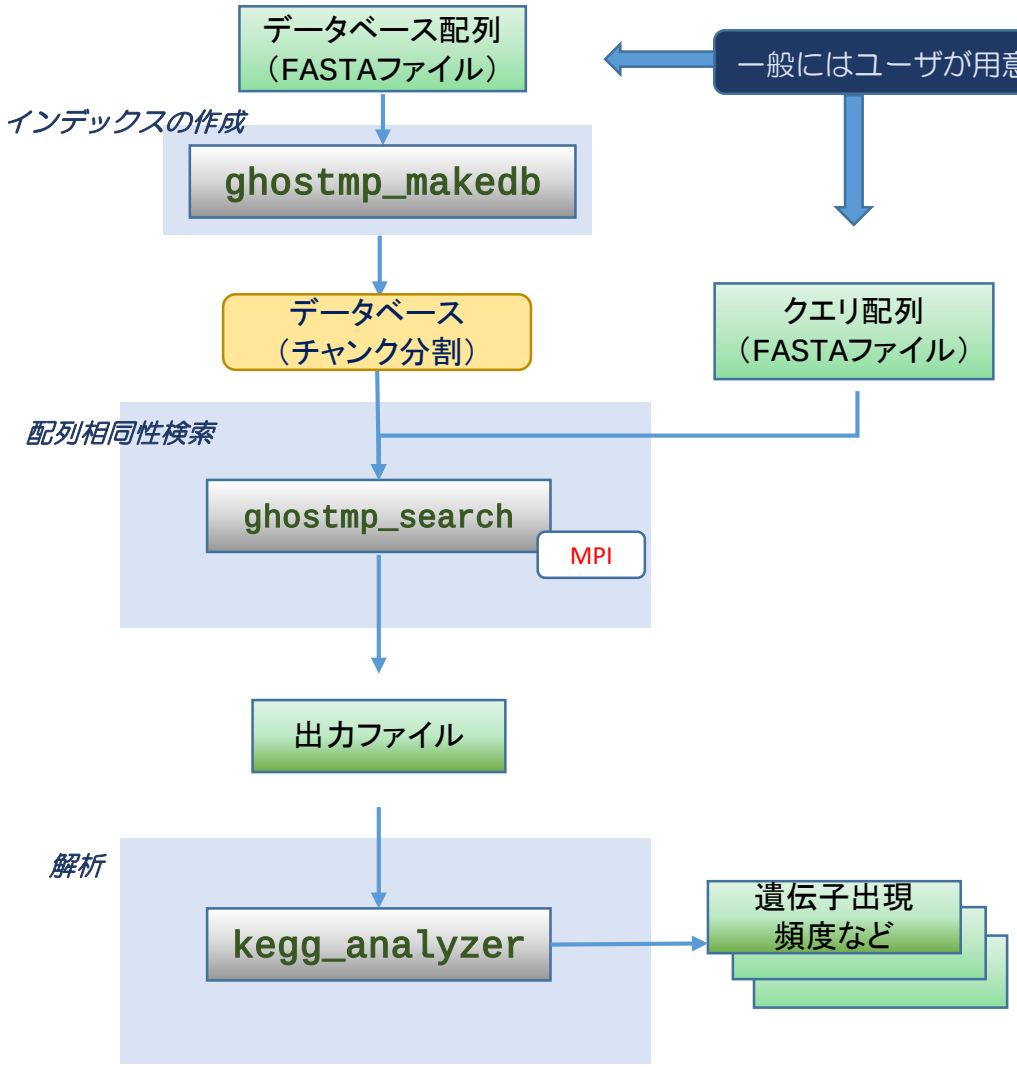


MPIライブラリによる効率的なファイルアクセス

実習のながれ

実習用にあらかじめ用意した クエリ配列 と データベース配列 を使って、
配列相同性検索の並列分散処理を実行し、検索結果の解析を行う。

GHOST-MP コマンドの実行のながれ



※ データベースが変わらない限り、インデックスの作成は繰り返す必要はありません。同じデータベースに対して検索する際は、以前に作成したものを使用できます。

※ 緑色 は テキストファイル
黄色 は バイナリファイル

アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

SCLS計算機システムへのログイン（1）

- sshによるログイン
  - ログインノード：hpci-scls.riken.jp
  - 事前に用意したsshの秘密鍵を使って、仮想端末などからログインする
    - Windows：　PuTTYやTera Term
    - LinuxやMac OS X：　sshコマンド

SCLS計算機システムへのログイン（2）

```
$ ssh username@hpci.scls.riken.jp
SCLS System Info: Service is available.
Last login: Thu Mar 10 14:34:45 2015 from foo.ac.jp
#------------------------------------------------------------------------------#
#   SCLS System Information                              Date : Mar. 9, 2015 #
#------------------------------------------------------------------------------#
#                                                                              #
#   Welcome to SCLS System                                                     #
#                                                                              #
#   If you have any questions or need for further assistance, please send a    #
#   message to scls-adm@riken.jp                                               #
#                                                                              #
#   For SCLS information, please refer to https://hpci-scls.riken.jp/          #
#                                                                              #
#------------------------------------------------------------------------------#
[username@scls ~]$
```

＊仮想端末の表示は環境設定等のため、必ずしも一致しない

アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

実習の準備

実習に必要なファイルのコピーと展開

- SCLS計算機システム上の
  「/home/kakuta/ghostmp_koshukai.tar.gz」をコピーする
- 必要ならば適宜ディレクトリを作成して、そのディレクトリの中で作業する

```
[username@scls ~]$ cp /home/kakuta/ghostmp_koushukai.tar.gz .

[username@scls ~]$ tar zxf ghostmp_koushukai.tar.gz

[username@scls ~]$ cd ghostmp_koushukai

[username@scls ~]$ ls
data  ghostmp-1.3.3  precomp  sample  script  workspace
```

実習用ファイルをコピー

実習用ファイルを解凍展開します

カレントディレクトリを変更

カレントディレクトリの内容を表示

ファイルの内訳

- data/ :
  実習に用いるメタゲノムデータ
- ghostmp-1.3.3/ :
  GHOST-MPのソースコード
- precomp/:
  事前に計算されたGHOST-MPの結果（時間がかかるため）
- sample/:
  実習で作成するスクリプトなどのサンプル
- script/ :
  GHOST-MPの出力結果の解析に用いるスクリプト
- workspace/:
  このディレクトリで実習を行うことを想定

アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

SCLS計算機システムの利用方法

ログイン（ssh）

ログインノード
(hpci-scls.riken.jp)

ジョブの投入
（pjsubなど）

計算ノード

計算ノード01　計算ノード02

計算ノード03　計算ノード04

⋮

＊ログインノードで高負荷の計算を行わない
ログインノードに負荷がかかると、
他のユーザーの利用に支障が出るため。

ジョブの実行方法（1）

- SCLS計算機システムでのジョブの実行には、キューイングシステムを利用する
- 次のように実行するジョブを記述したスクリプトを作成し、
  pjsubコマンドでジョブのサブミットを行う

```
[username@scls workspace]$ pwd
/home/username/ghostmp_koushukai/workspace
[username@scls workspace]$ mkdir job_submit
[username@scls workspace]$ cd job_submit
[username@scls job_submit]$ vim run.sh
```

スクリプトの例

run.sh

```
#!/bin/sh
#PJM -L “rscgrp=small“
#PJM -L “node=1”
#PJM -L “elapse=00:10:00”

sleep 120
echo “test message”
```

- リソースグループスモールを指定
- 計算ノード、１ノードを使用
- 最大経過時間10分
- 120秒間待ってから
- “test message”と標準出力に出力する

ジョブの実行方法（2）

- キューイングシステムの主なコマンド
  - pjsub: ジョブの投入
    - pjsub sample_job.sh
  - pjstat : 投入されたジョブの状態表示
    - pjstat
  - pjdel : 投入されたジョブの削除
    - pjdel job_id

ジョブの実行方法（3）

run.shの投入例

```
[username@scls]$ pjsub run.sh
[INFO] PJM 0000 pjsub Job 19247 submitted.
[username@scls]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1
s      0      0      0      0      1      0      0      0      0       1

JOB_ID     JOB_NAME   MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
19247      run.sh     NM RUN username 03/11 11:42:58  0000:05:00 1
[usename@scls]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      0      0      0      0      0       0
s      0      0      0      0      0      0      0      0      0       0

[username@scls]$ ls
run.sh  run.sh.e19247  run.sh.o19247
[username@scls]$ cat run.sh.o19247
test message
[username@scls]$
```

作成したスクリプトを投入する

投入したジョブの状態を出力する

ジョブが終了するとpjsubで表示されなくなる

サブミット後にジョブの標準出力と標準エラー出力がrun.sh.o<job_id>とrun.sh.e<job_id>に出力される

ジョブの状態（1）

```
[username@scls]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1        [1]
s      0      0      0      0      1      0      0      0      0       1        [2]

JOB_ID      JOB_NAME    MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
19247       run.sh      NM RUN username 03/11 11:42:58  0000:05:00 1             [3]
```

[1] ある状態のジョブ数
[2] サブジョブを考慮したジョブ数
　（本実習ではサブジョブは使用しない）
[3] 各ジョブの情報

[1]および[2]	内容
ACCEPT	受入中のジョブ数
QUEUED	ステージイン待機中のジョブ数
STGIN	ステージイン中のジョブ数
READY	実行待機中のジョブ数
RUNING	実行中のジョブ数
RUNOUT	ジョブ終了中のジョブ数
STGOUT	ステージアウト中のジョブ数
HOLD	ユーザー指示で待機中のジョブ数
ERROR	エラーのため待機中のジョブ数
TOTAL	全ジョブ数

ジョブの状態（2）

```
[username@scls]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1        [1]
s      0      0      0      0      1      0      0      0      0       1        [2]

JOB_ID     JOB_NAME   MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
19247      run.sh     NM RUN username 03/11 11:42:58  0000:05:00 1             [3]
```

[3]	内容
JOB_ID	キューイングシステムによってジョブに割り振られたID
JOB_NAME	ジョブの名前。ユーザーが指定できる。未指定の場合は、スクリプト名。
MD	ジョブの種類（通常のジョブ、ステップジョブ、バルクジョブ）
ST	ジョブの状態
USER	ジョブを投入したユーザー
START_DATE	ジョブの実行開始日時
ELAPSE_LIM	ジョブがRUN状態になってからのの経過時間
NODE_REQUIRE	要求ノード

ST	内容
ACC	受入中
RJT	棄却中
QUE	実行待機中
SIN	ステージイン中
RDY	実行待機中（SIN後）
RNA	計算資源確保中
RUN	実行中
RNO	ステージアウト待機中
SOT	ステージアウト中
EXT	ジョブ終了処理完了
CCL	ジョブ実行中止
HLD	ユーザーの指示で待機中
ERR	エラー状態

ジョブの削除

- 誤ったジョブを投入してしまったときなどは、ジョブの削除が可能
  - pjdel：投入されたジョブの削除
    - pjdel job_id
    - 削除したいジョブのJOB_IDをpjstatなどで調べて、pjdelコマンドの引数として指定する
    - 待機中のジョブも、実行中のジョブの削除可能

```
[username@scls]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1
s      0      0      0      0      1      0      0      0      0       1

JOB_ID      JOB_NAME   MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
19247       run.sh     NM RUN username 03/11 11:42:58  0000:05:00 1
[username@scls]$ pjdel 19247
```

アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

アジェンダ

- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）

STEP 3-3までの作業で得られる結果
　各クエリ配列の相同配列とそのアラインメント情報

```
query_name  sbj_name  0.714286  14  2  1  32  73  105  116  2.88961  29.4819
```

各カラムが示すもの

```
#query_name sbj_name identity(%) alignment_length num_mismatch num_gap query_start
query_end sbj_start sbj_end e_value bit_score
```

コンパイル

Requirement

- OpenMPに対応したコンパイラ
- MPI Library
- tr1またはBoost C++ Library

ghostmp_makedb **ghostmp_search**

```
$ cd ~/ghostmp_koushukai/ghostmp-1.3.3/src/
$ make
    ⋮

$ ls ghostmp_makedb ghostmp_search
ghostmp_search ghostmp_makedb
```

コンパイルに成功すると ghostmp_makedb と ghostmp_search が作成される

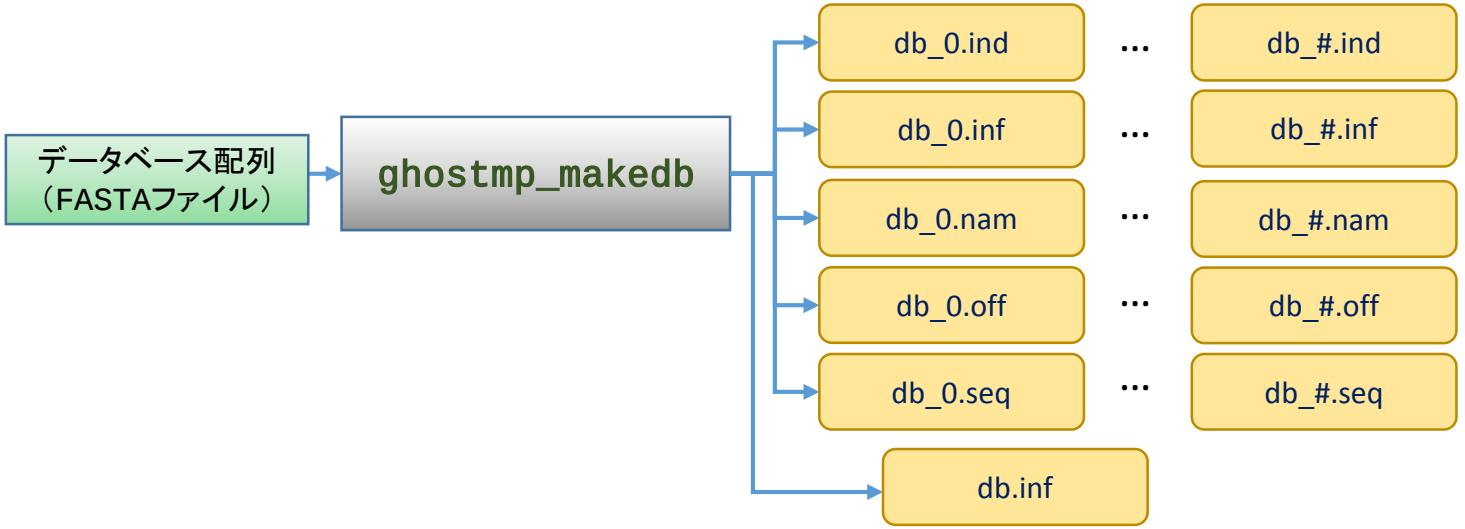
アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

扱うクエリ、データベースについて

- 注意
  - 計算時間短縮のため、「データベースのインデックス作成」と「配列相同性検索」では、実際に解析に用いられるファイルの一部が用いられています。
  - 「検索結果の集計」からは、事前計算された結果を利用して頂きますので、ファイルを取り違えないように注意して下さい。

インデックスの作成［ghostmp_makedb］（入力ファイル）



配列相同性検索の参照先となる FASTA フォーマットのアミノ酸配列の
データベースファイルからインデックスを作成する

FASTA フォーマットとは

```
>gi|66816243|ref|XP_642131.1| hypothetical protein
MASTQNIVEEVQKMLDTYDTNKDGEITKAEAVEYFKGKKAFNPERSAIYLFQVYDKDNDGKITIKELAGDIDFDKALKEY
KEKQAKSKQQEAEVEEDIEAFILRHNKDDNTDITKDELIQGFKETGAKDPEKSANFILTEMDTNKDGTITVKELRVYYQK
VQKLLNPDQ
```

“>”で始まるヘッダ行　配列の説明　1行で書く

配列データ　アミノ酸を一文字表記で表している　改行 OK

データベースサイズが大きい場合（数GB～数十GB）#個 のチャンクに分割して扱う。
それぞれのチャンクについて 5つのファイル (.ind .inf .nam .off .seq) が作成される。

インデックスの作成［ghostmp_makedb］（実行準備）

ghostmp_makedb

ghostmp_makedb -i INPUT -o OUTPUT
- INPUT: データベース配列
- OUTUPT: 作成するインデックスのファイル名

ジョブスクリプト（逐次ジョブ）の作成

```
[username@scls job_submit]$ pwd
/home/kakuta/ghostmp_koushukai/workspace/job_submit
[username@scls job_submit]$ mkdir ../makedb && cd $_
[username@scls makedb]$ vim makedb.sh
```

makedb.sh

```
#!/bin/sh
#------ pjsub options ------#
#PJM -L "rscgrp=small"
#PJM -L "node=1"
#PJM -L "elapse=00:10:00"
#---------------------------#
BIN="${HOME}/ghostmp_koushukai/ghostmp-1.3.3/src/ghostmp_makedb"
INPUT="${HOME}/ghostmp_koushukai/data/kegg.genes.pep.201106.reduced"
OUTDIR="out"
OUTPUT="${OUTDIR}/db"
ARGS="-i ${INPUT} -o ${OUTPUT}"

mkdir ${OUTDIR}
${BIN} ${ARGS}
```

リソースグループ「small」
使用ノード数「1」
最大経過時間「10分間」

ghostmp_makedb プログラムのパス

データベース配列（KEGG GENES アミノ酸配列の一部）

ghostmp_makedb の実行

インデックスの作成［ghostmp_makedb］（ジョブの投入）

ジョブの投入 と 状態確認

```
[username@scls makedb]$ pjsub makedb.sh
[INFO] PJM 0000 pjsub Job 10569 submitted.
[username@scls makedb]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1
s      0      0      0      0      1      0      0      0      0       1

JOB_ID     JOB_NAME   MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
10569      makedb.sh  NM RUN username 10/24 14:49:45  0000:10:00 1
```

ジョブ実行結果

```
[username@scls makedb]$ cat makedb.sh.o*


The number of chunks :1
Max length of a chunk : 50476865
Total database length : 50369581
Total number of sequences : 107283
```

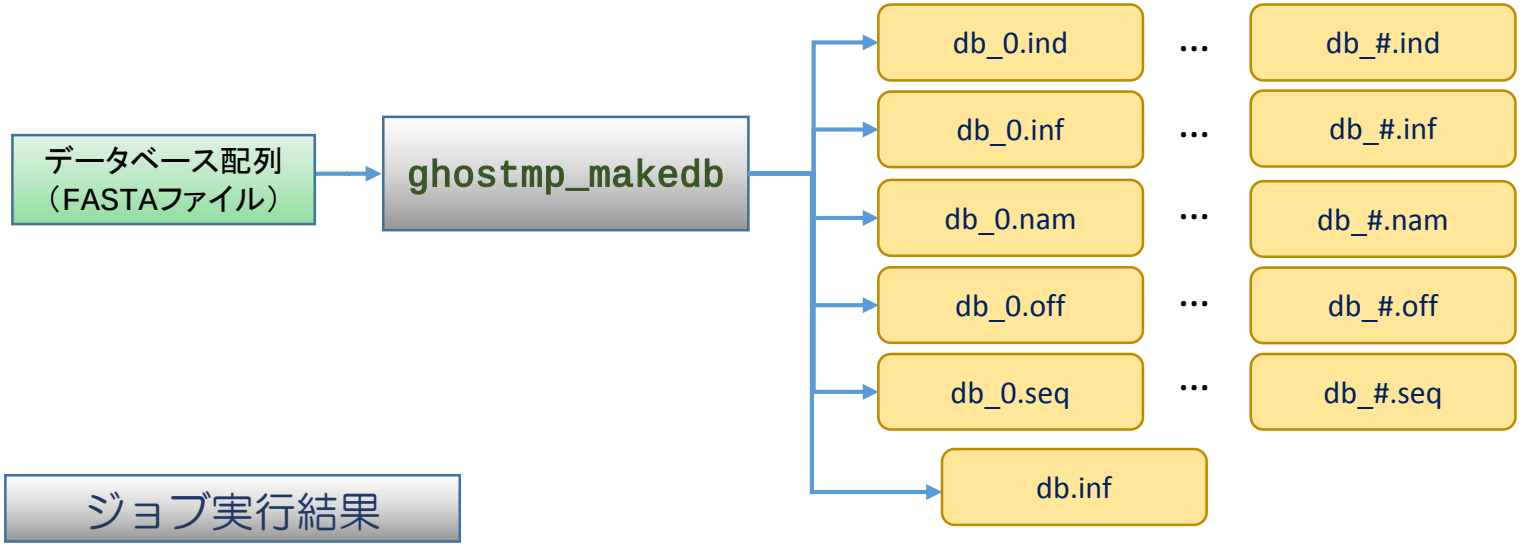
インデックスのチャンク数1

チャンクの最大長

配列の合計長

全配列数

インデックスの作成［ghostmp_makedb］（出力ファイルの確認）



ジョブ実行結果

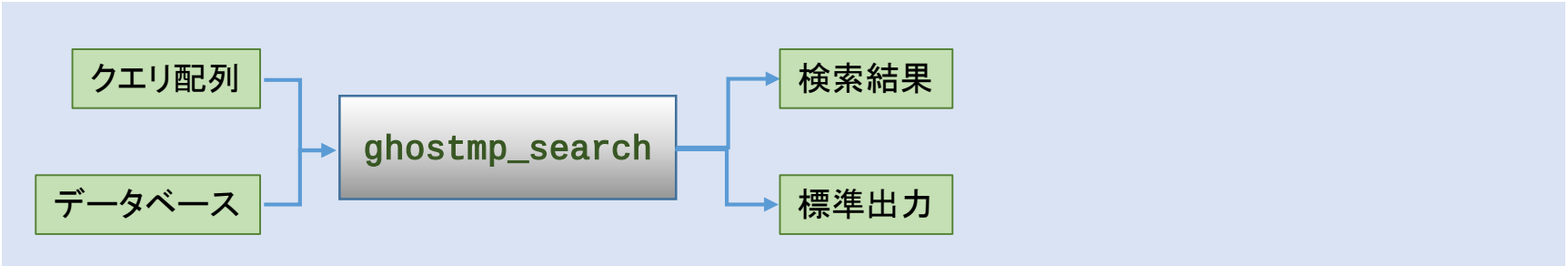
```
[username@scls makedb]$ ls -l out
total 575200
-rw-r--r-- 1 username group 529479035 Oct 24 14:50 db_0.ind
-rw-r--r-- 1 username group         8 Oct 24 14:50 db_0.inf
-rw-r--r-- 1 username group   8468034 Oct 24 14:50 db_0.nam
-rw-r--r-- 1 username group    429132 Oct 24 14:50 db_0.off
-rw-r--r-- 1 username group  50476865 Oct 24 14:50 db_0.seq
-rw-r--r-- 1 username group        24 Oct 24 14:50 db.inf
```

チャンク0　データベース

アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

配列相同性検索［ghostmp_search］（入力ファイル）



入力ファイル確認 ― FASTAフォーマットの塩基配列

```
[username@scls workspace]$ less ../data/query
```

クエリファイル

```
>61JGYAAXX100510:7:100:10000:7489/1
AACCCAACATGGCATCCTTGGTCCCTAGAGCAATCTCCTTGCCCTCTTTTTTAGCATAACTGATTAACTGGCGCAAATGTAACTTGG
AAATTGGGCTCGTG
>61JGYAAXX100510:7:100:17664:14710/1
ACTTGCAATCTGCGATTCTGTTTGCCCAGACTCCGACAAAATTTTCACCTGAGTAAAAGTCCGAAATTGTCCTAGCATGACTGCAAT
CAACTTGATTTCAT
  ⋮
```

ヘッダ行

塩基配列

※ クエリファイルの配列は、次世代シーケンサーで得られた　HMPの頬粘膜のDNA塩基配列データ。

配列相同性検索［ghostmp_search］（実行準備）

ghostmp_search

ghostmp_search -a NUM_THREADS -q QUERY_TYPE -i INPUT -d DB -o OUTPUT

- NUM_THREADS: スレッド数
- QUERY_TYPE: クエリが塩基配列ならば”d”、アミノ酸配列ならば”p”を指定する
- INPUT: クエリ配列
- DB: ghostmp_makedbで作成したインデックスのprefix
- OUTUPT: 検索結果の出力先

ジョブスクリプト（MPI並列ジョブ）

```
[username@scls workspace]$ mkdir search && cd $_
[username@scls search]$ vim search.sh
```

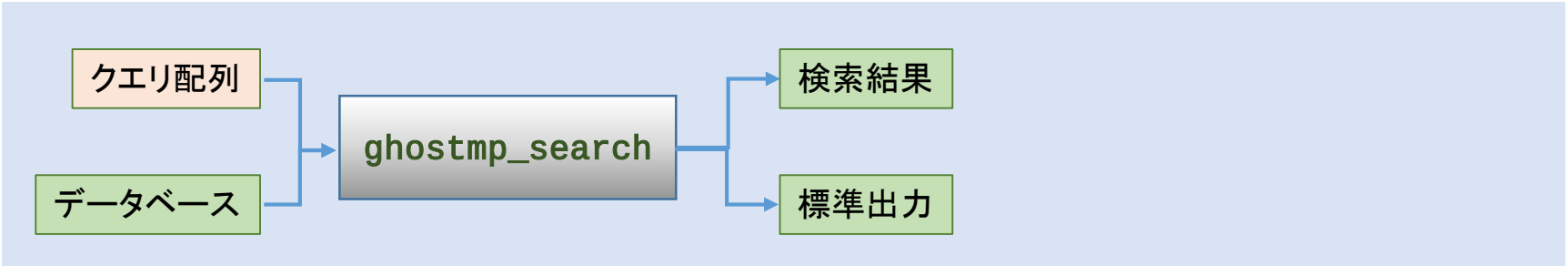
search.sh

```
#!/bin/sh
#------ pjsub options ------#
#PJM -L "rscgrp=small"
#PJM -L "node=2"
#PJM --mpi "proc=2"
#PJM -L "elapse=00:10:00"
#---------------------------#

export OMP_NUM_THREADS=16

mpiexec -n 2 ../../ghostmp-1.3.3/src/ghostmp_search -a ${OMP_NUM_THREADS} ¥
-q d -i ../../data/query -d ../makedb/out/db -o search_result
```

リソースグループ「small」
使用ノード数「2」
プロセス数「2」
最大経過時間「10分」

スレッド並列数の設定

配列相同性検索［ghostmp_search］（ジョブの投入）



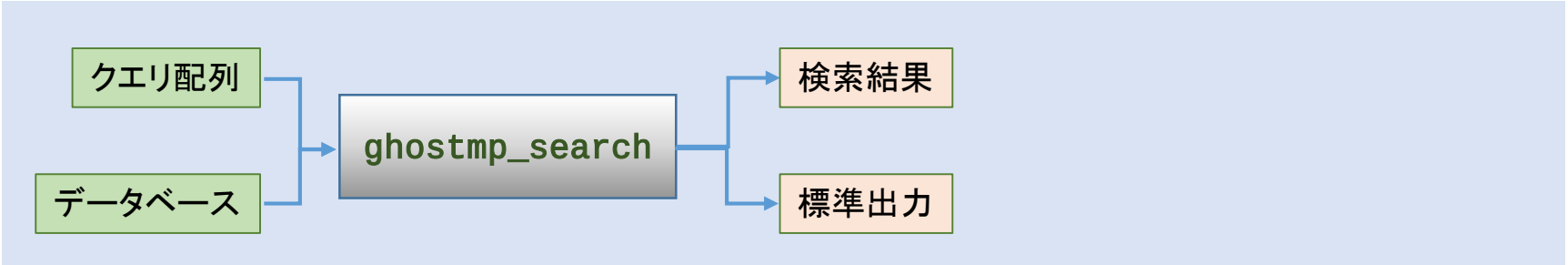
ジョブの投入と状態確認

```
[username@scls search]$ pjsub search.sh

[INFO] PJM 0000 pjsub Job 10583 submitted.
[username@scls search]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      1      0      0      0      0       1
s      0      0      0      0      1      0      0      0      0       1

JOB_ID     JOB_NAME   MD ST  USER     START_DATE      ELAPSE_LIM NODE_REQUIRE
10583      search.sh  NM RUN username 10/24 16:23:48  0000:10:00 2
```

ジョブ終了確認

```
[username@scls search]$ pjstat

  ACCEPT QUEUED  STGIN  READY RUNING RUNOUT STGOUT   HOLD  ERROR   TOTAL
       0      0      0      0      0      0      0      0      0       0
s      0      0      0      0      0      0      0      0      0       0
```

配列相同性検索［ghostmp_search］（出力ファイルの確認）



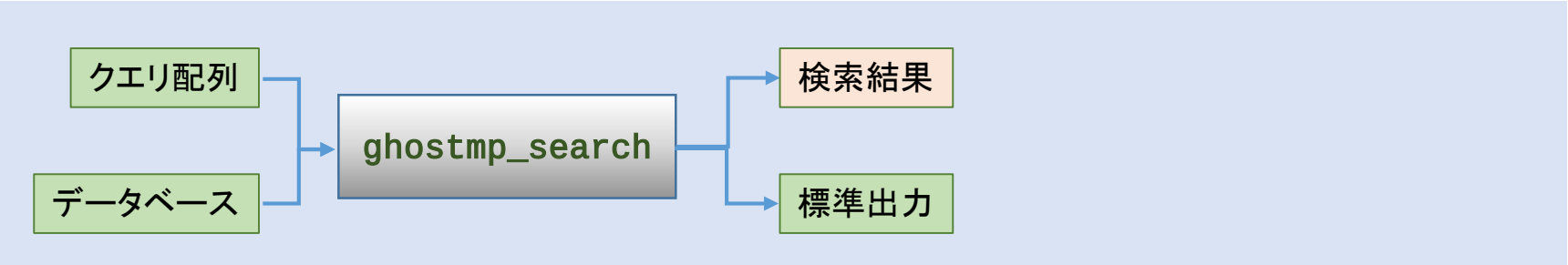
結果確認1 ― 出力ファイル

```
[username@scls search]$ ls -l

total 2880
-rw-r--r-- 1 username group 4455377 Oct 24 16:24 search_result
-rw-r--r-- 1 username group     324 Oct 24 16:20 search.sh
-rw-r--r-- 1 username group       0 Oct 24 16:24 search.sh.e10583
-rw-r--r-- 1 username group       0 Oct 24 16:24 search.sh.o10583
```

ghostmp_searchの検索結果

ghostmp_searchの標準出力

配列相同性検索［ghostmp_search］（出力内容の確認）



結果確認1 ー 出力ファイル

```
[username@scls search]$ less search_result

61JGYAAXX100510:7:100:10000:7489/1      mmu:12406  Serpinh1, BERF-1, Cbp1, Cbp2, Hsp47,
J6, Serpinh2, colligin, gp46; serine (or cysteine) peptidase inhibitor, clade H, member
1; K09501 serpin peptidase inhibitor, clade H, member 1 50      4       2       0
67      99      121     131     105.573 23.483
   ⋮
```

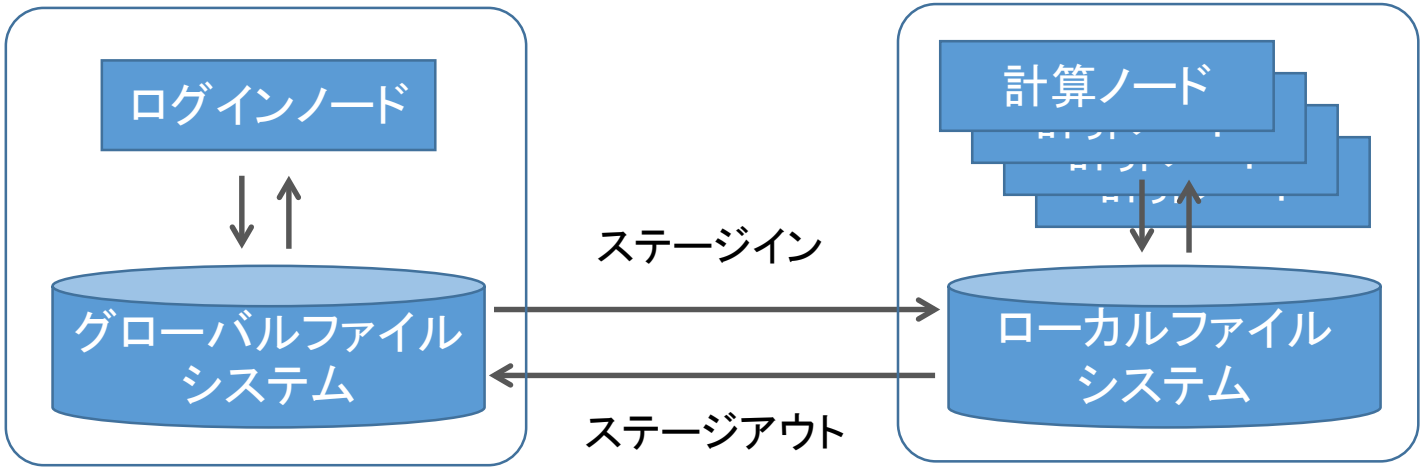
ghostmp_searchの検索結果

※検索結果は、1行ごとにTAB区切りで出力される

クエリ配列名、ヒット配列名、一致率、アラインメント長、不一致数、ギャップ数、
クエリ配列におけるアラインメント開始位置、その終了位置、
ヒット配列におけるアラインメント開始位置、その終了位置、
E-value、bitスコア

「京」での実行スクリプト例

「京」とSCLS計算機システムは互換性の高いシステムだが、
「京」ではジョブに必要なファイルを転送するための記述が必要



「京」での実行スクリプト例

```
#!/bin/sh
#PJM --L "rscgrp=small"
#PJM --L "elapse=0:10:00"
#PJM --L "node=2"
#PJM --mpi "use-rankdir"
#PJM --stgin “rank=* ../../ghostmp-1.3.3/src/ghostmp_search %r:./ghostmp_search”
#PJM --stgin “rank=0 ../../data/query 0:../”
#PJM --stgin “rank=0 ../makedb/out/db* 0:../db/”
#PJM --stgout “rank=0 0:../search_result ./search_result”
...
```

ステージインとステージアウトに関する記述

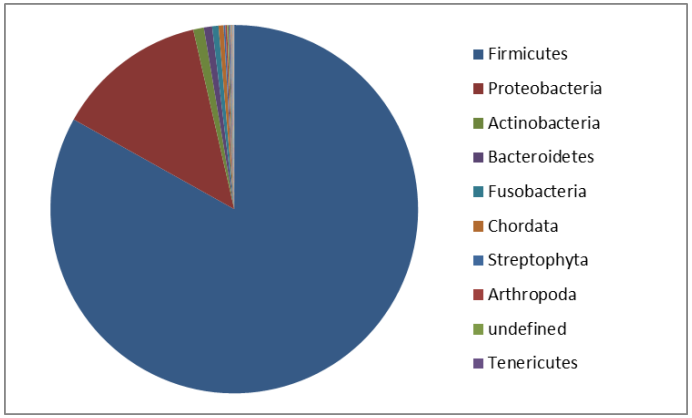
アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

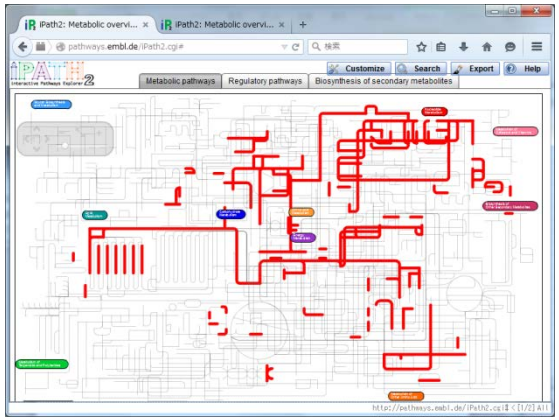
アジェンダ

- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

STEP 4-3までの作業で得られる結果



STEP4-2 : サンプル内のphylumの割合



STEP4-3 : サンプル内の遺伝子が関わる代謝パスウェイの可視化

検索結果の集計（1）

GHOST-MP の出力ファイルを一個にまとめる

```
[username@scls search]$ mkdir ../summary && cd $_
[username@scls summary]$ ls ../../precomp/
result.buccal_mucosa  result.subgingival_plaque
```

result.buccal_mucosa: ヒトの頬粘膜（HMPのSRS011090）の検索結果
result.supragingival_plaque: ヒトの歯肉縁上の歯垢（HMPのSRS044366）の検索結果
（HMP: Human Microbiome Project。ヒトの体表・体内のメタゲノム研究プロジェクト）

これらファイルは、実習で作成したものではなく、事前の計算結果を配布してあるもの

解析を実行する

```
[username@scls summary]$ mkdir buccal_mucosa && cd $_
[username@scls buccal_mucosa]$ python ../../../script/summarize_search_result.py ¥
../../../precomp/result.buccal_mucosa
```

集計用スクリプトの実行　事前計算の結果を引数にする

supragingival_plaqueでも同様に解析を実行する

検索結果の集計（2）

解析を実行する　標準出力

```
[user1@scls summary]$ python ../../script/summarize_search_result.py ¥
../../precomp/result.buccal_mucosa

2013/11/21 14:44:21     START   Start KEGG Analyzer
2013/11/21 14:44:21     START   Loading Gi-Taxid map file...
2013/11/21 14:45:27     END     54380376 genes loaded.
2013/11/21 14:45:27     START   Loading Taxonomy root map file...
2013/11/21 14:45:33     END     919194 species loaded.
2013/11/21 14:45:33     START   Loading KO-Enzyme map file...
2013/11/21 14:45:33     END     4808 KOs loaded.
2013/11/21 14:45:33     START   Loading USCG map file...
2013/11/21 14:45:33     END     36 USCGs loaded.
2013/11/21 14:45:33     START   Loading KEGG genes file...
2013/11/21 14:46:22     END     8782317 genes loaded.
2013/11/21 14:46:24     START   Loading Blast result...
2013/11/21 14:46:25     END     15000 blast results loaded.
2013/11/21 14:46:25     START   Normalizing...
2013/11/21 14:46:25     START   1. count genes...
2013/11/21 14:46:26     END     done.
2013/11/21 14:46:26     START   2. count USCGs...
2013/11/21 14:46:26     END     done.
2013/11/21 14:46:26     START   3. normalizing...
2013/11/21 14:46:28     END     done.
2013/11/21 14:46:28     END     Normalize.
2013/11/21 14:46:28     END     End KEGG Analyzer
```

検索結果の集計（3）

出力ファイル（主なもの）

ファイル名	内容
<search_result>.genes_freq	遺伝子 の出現頻度
<search_result>.ko_ratio	各KO（KEGG Orthology）の割合
<search_result>.phylum_ratio	phylum の割合

集計方法

- 検索結果全体から、各遺伝子の出現頻度を推定する。
- 各遺伝子に対するKEGGのアノテーションからKOについて集計する。
- pylumとgenusについてはマーカー遺伝子のみを用いて推定を行う。

アジェンダ

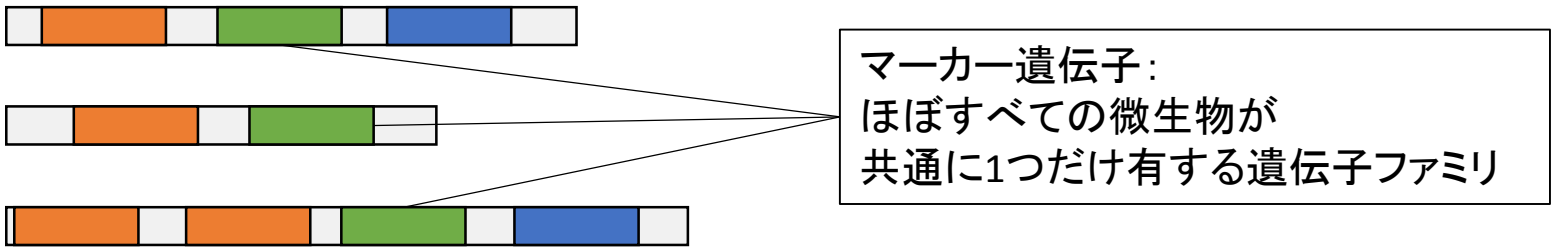
- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

分類群の相対存在度推定のためのマーカー遺伝子の利用

von Mering, C. et al. Science 2007, 31, 1126-1130

微生物によってゲノムサイズや有する遺伝子ファミリ、その遺伝子ファミリのコピー数や遺伝子長などは異なる。
このため、そのまま検索結果に表れた回数などから、相対存在度を推定することはできない。

ほぼ全ての微生物が共通に1つだけ有する遺伝子ファミリをマーカーとして、それらが検索結果に表れた回数から遺伝子長の補正を加えて相対存在度の推定を行う



その他、分岐群特異的なマーカー遺伝子を用いる方法もある

分類群の相対存在度（１）

phylum_ratio ファイルの確認

マーカー遺伝子からpylum ratioを推定

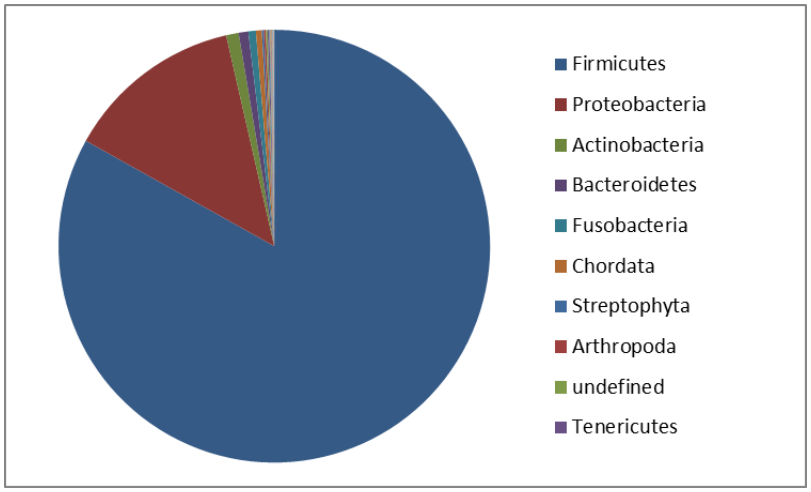
```
[username@scls buccal_mucosa]$ sort -k 2 -gr result.buccal_mucosa.phylum_ratio | less

Firmicutes      0.830939
Proteobacteria  0.133185
Actinobacteria  0.009338
Bacteroidetes   0.007251
Fusobacteria    0.005592
Chordata        0.004407
Streptophyta    0.001515
Arthropoda      0.001482
undefined       0.001363
Tenericutes     0.001238
Apicomplexa     0.000567
Cyanobacteria   0.000496
Spirochaetes    0.000365
Ascomycota      0.000346
Platyhelminthes 0.000279
Euryarchaeota   0.000254
Nematoda        0.000230
Chlamydiae      0.000169
Chlorobi        0.000160
Thermotogae     0.000148
```

pylumを割合の降順に表示

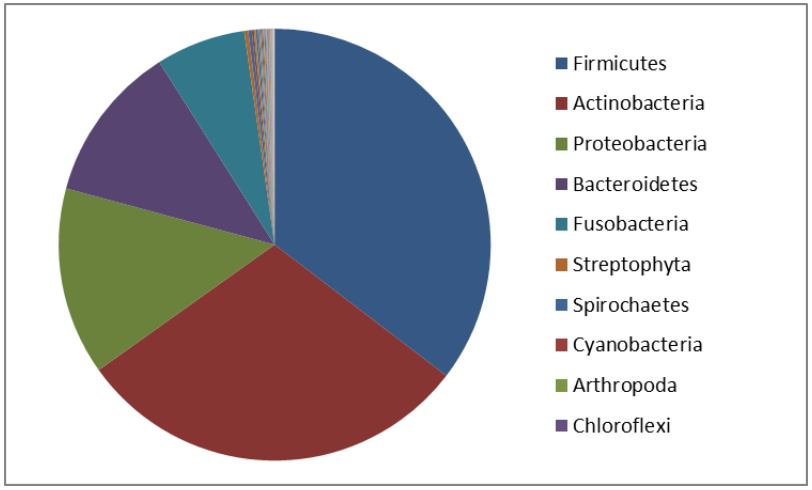
分類群の相対存在度（2）

phylum_ratioの数値を利用して
Excelなどで作図する

頬粘膜



歯肉縁上の歯垢



アジェンダ

- GHOST-MP実行のながれ
- STEP 1: 実習の準備
  - STEP 1-1: SCLS計算機システムへのログイン
  - STEP 1-2: 必要なファイルのコピーと展開
- STEP 2: ジョブスケジューラの利用方法
- STEP 3: GHOST-MPによる配列相同性検索
  - STEP 3-1: GHOST-MPのコンパイル
  - STEP 3-2: データベースのインデックス作成（ghostmp_makedb）
  - STEP 3-3: 配列相同性検索（ghostmp_search）
- STEP 4: メタゲノムデータに対する相同性検索結果の解析
  - STEP 4-1: 検索結果の集計
  - STEP 4-2: 分類群に基づく解析
  - STEP 4-3: 遺伝子オーソログに基づく解析

遺伝子オーソログの相対存在度

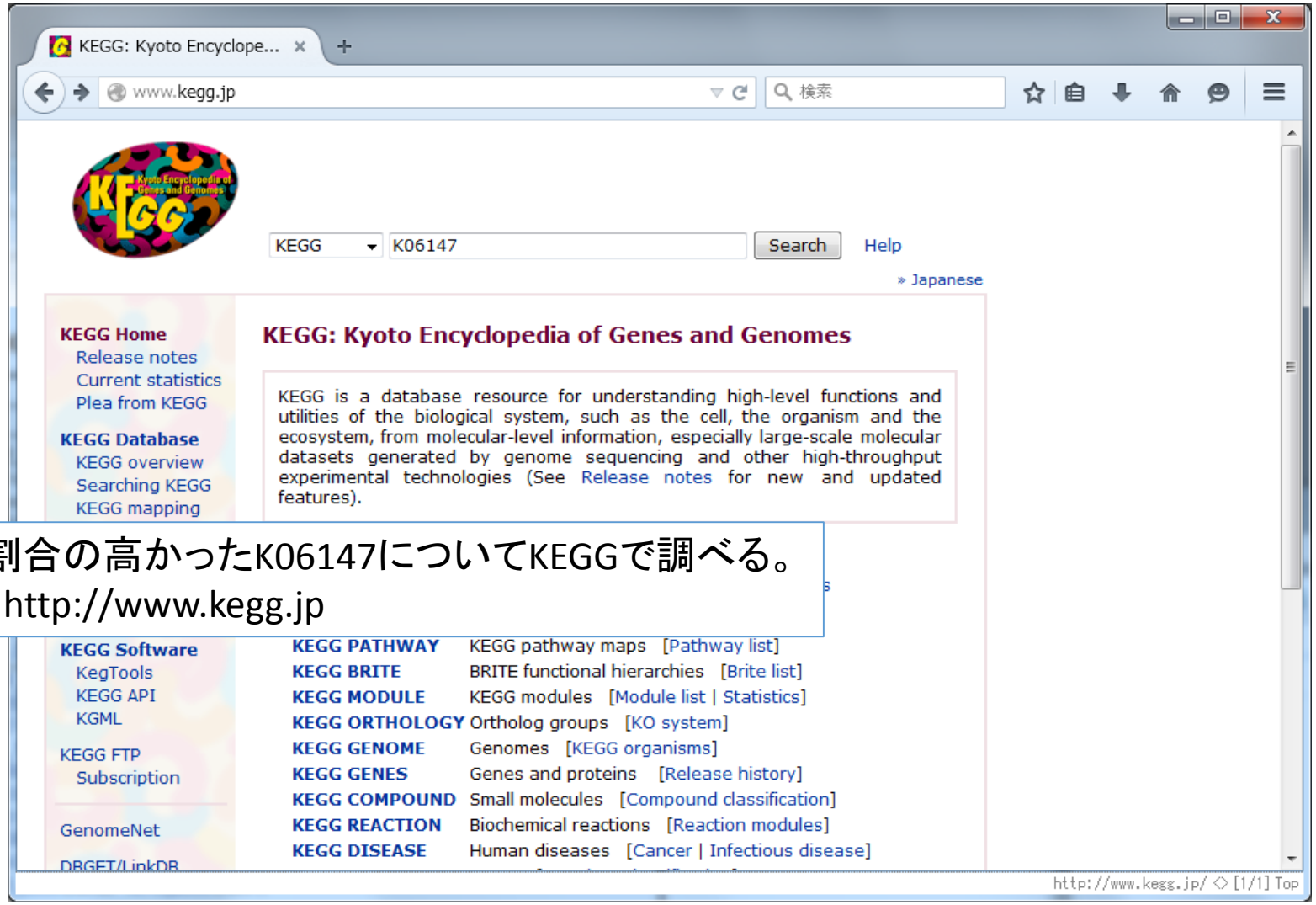
ko_ratio ファイルの確認

```
[username@scls buccal_mucosa]$ sort -k 2 -gr result.buccal_mucosa.ko_ratio | less
```

koを割合の降順に表示

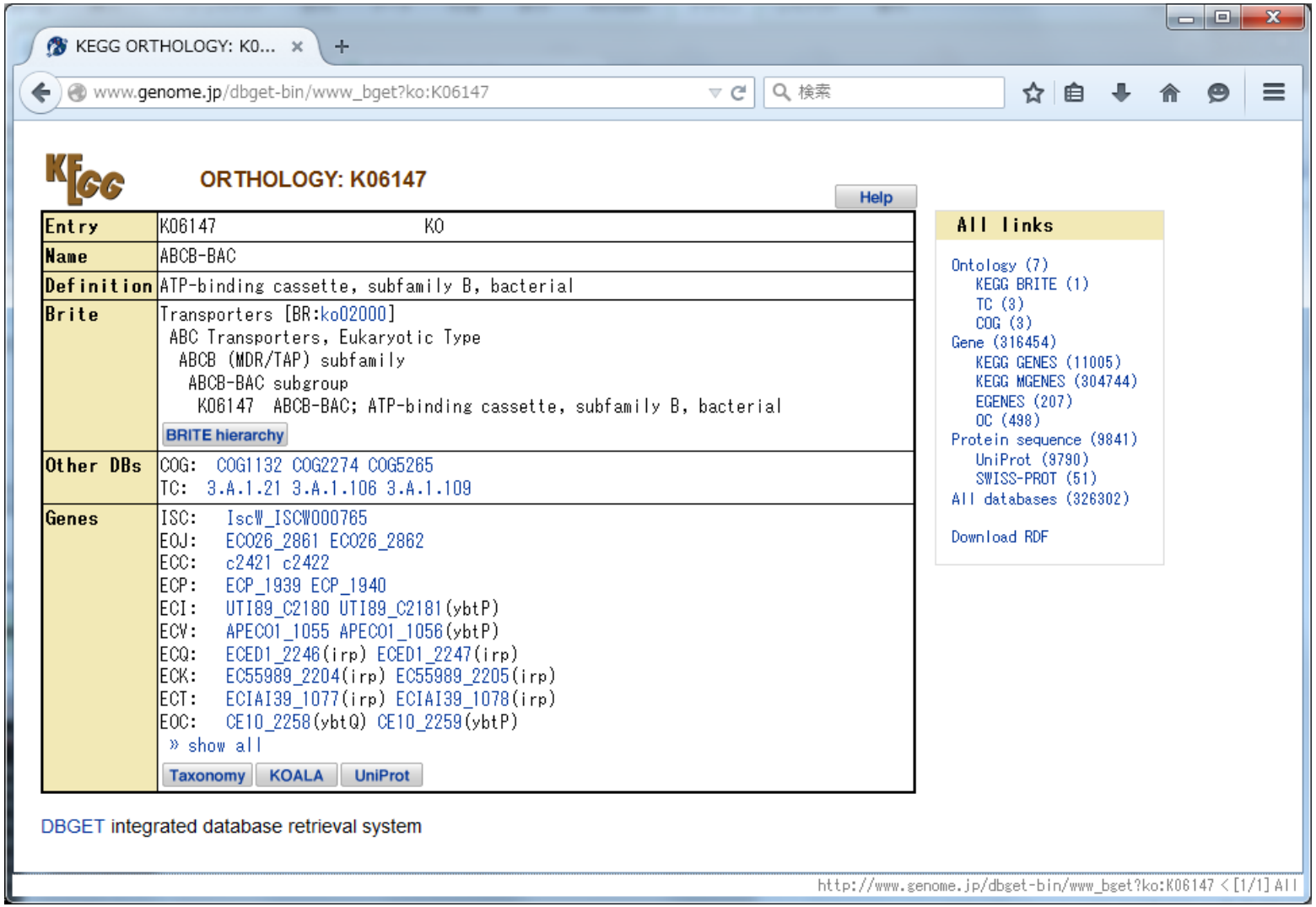
```
ko:K06147       0.003332
ko:K02029       0.003305
ko:K07052       0.003097
ko:K07024       0.002670
ko:K02003       0.002307
ko:K01992       0.002025
ko:K02028       0.002018
ko:K02006       0.001965
ko:K02030       0.001963
ko:K02004       0.001912
ko:K09687       0.001795
ko:K02529       0.001585
ko:K02913       0.001574
ko:K03574       0.001541
ko:K03402       0.001358
ko:K02078       0.001343
ko:K02008       0.001321
ko:K02015       0.001293
ko:K02793       0.001285
ko:K01462       0.001285
```

WebブラウザによるKEGG Orthologyの閲覧（1）



最も割合の高かったK06147についてKEGGで調べる。
URL：http://www.kegg.jp

WebブラウザによるKEGG Orthologyの閲覧（2）



http://www.genome.jp/

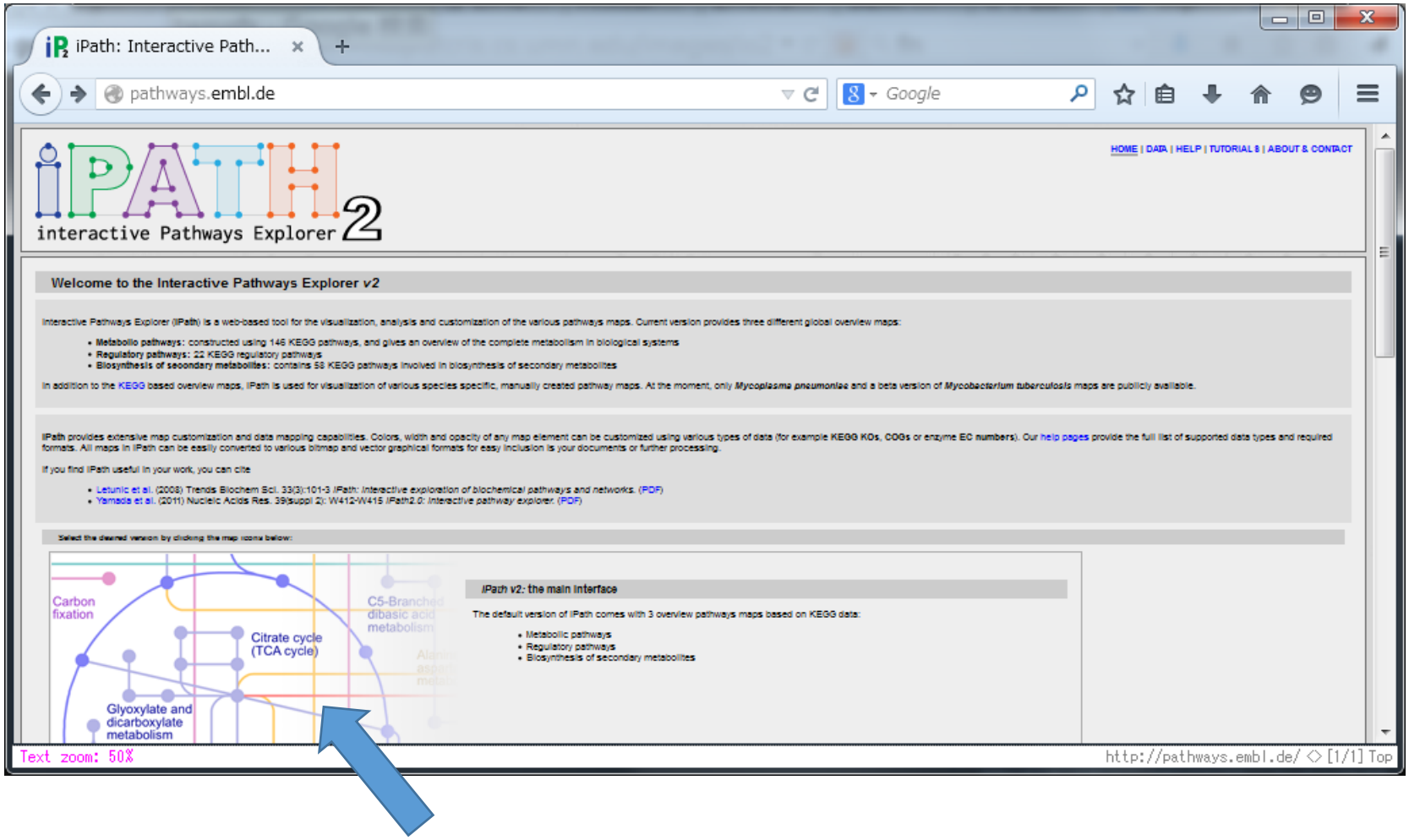
iPATH2による遺伝子オーソログ関連パスウェイの可視化

ここではiPATH2を用いて、検索結果に表れた
遺伝子オーソログが関連しているパスウェイを可視化する。
多くの遺伝子に関する情報を直観的に捉えることができる。

http://pathways.embl.de/

Yamada, T., et al. *Nucleic Acids Research* 2011, 39 (suppl 2), W412-W415

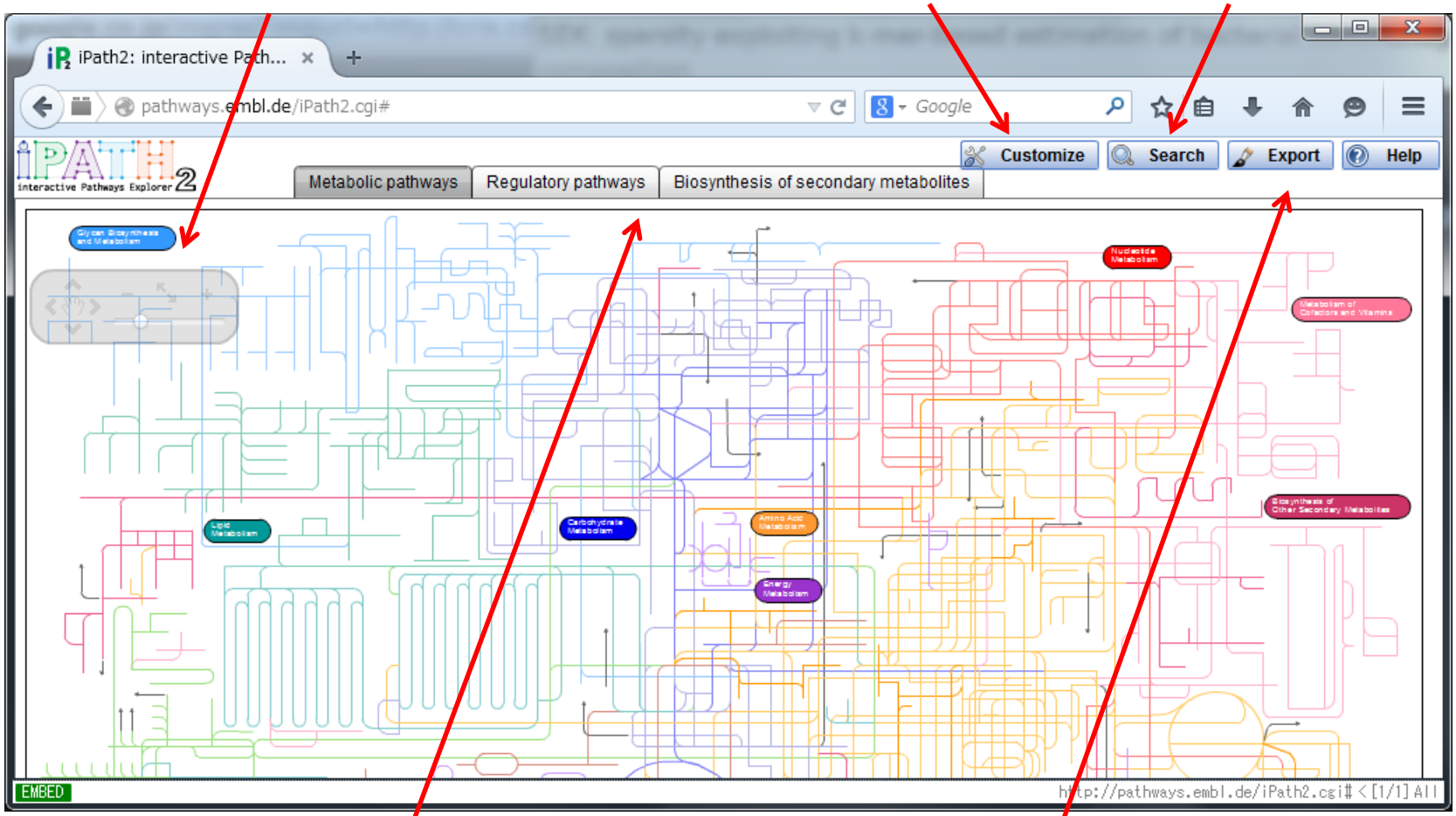
iPATH2によるパスウェイの可視化



http://pathways.embl.de/にアクセスして、ここをクリック

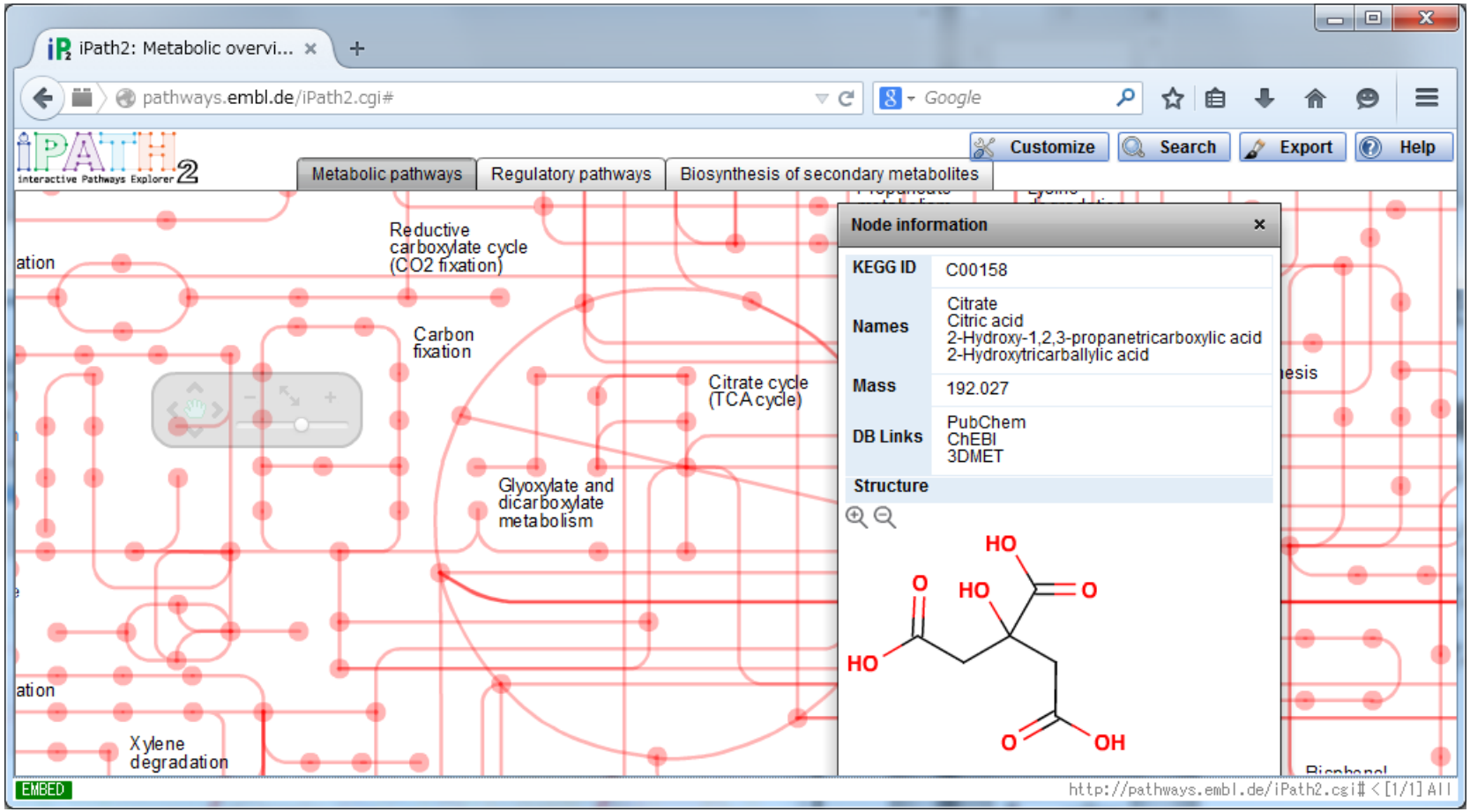
iPATH2のインターフェース

パスウェイの拡大・縮小　　表示するパスウェイの編集　　検索



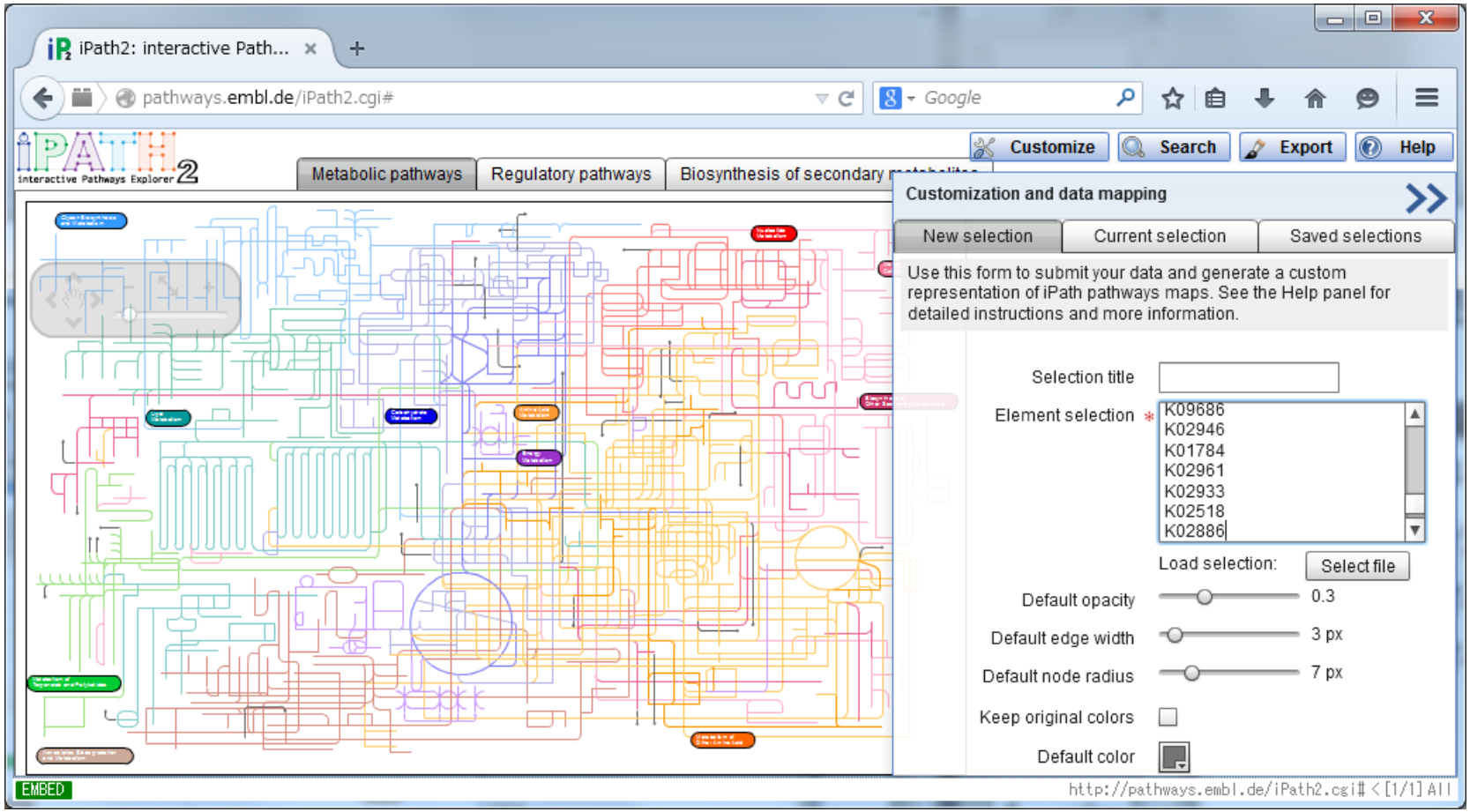
表示するパスウェイの選択　　画像として出力

パスウェイの要素を表示



ノードやエッジをクリックすると、化合物や反応に関する情報を表示することができる

iPATH2の表示のカスタマイズ



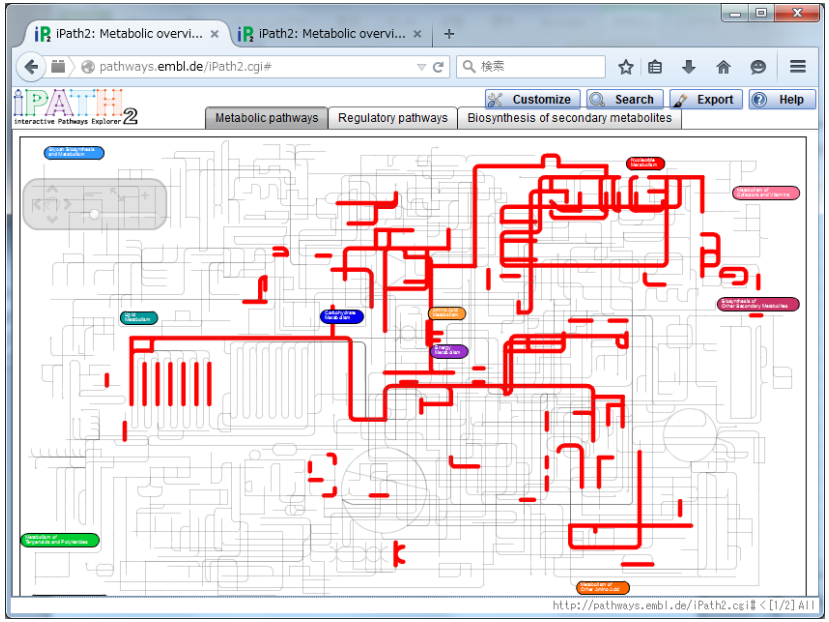
「Customize」ボタンでパネルを表示してElement selectionを指定すると対応するパスウェイを強調表示できる。KEGG KOの例

iPATH2を用いたサンプル間の比較

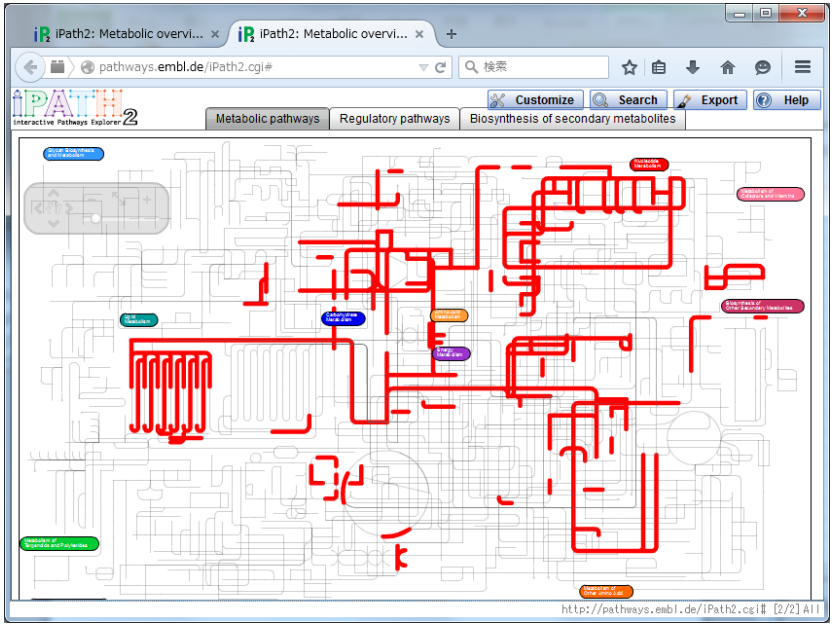
類粘膜と歯肉縁上の歯垢の結果をiPATH2を用いて比較する

例：割合が0.05%以上のKOの関わる代謝パスウェイをハイライトした

類粘膜

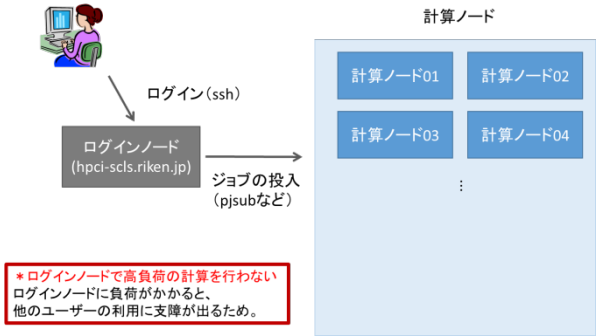


歯肉縁上の歯垢

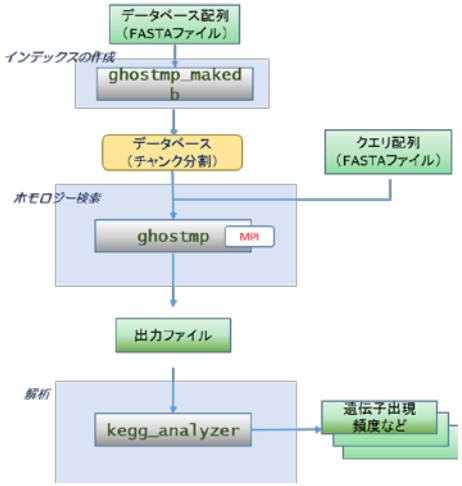


本日の実習内容

SCLS計算機システム上でのGHOST-MPの利用

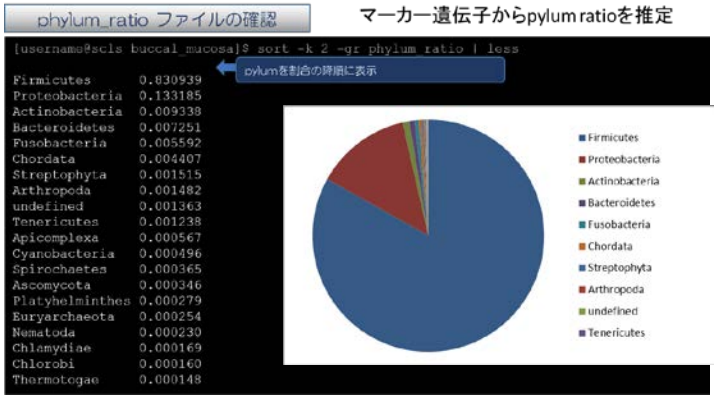


SCLS計算機システムの利用方法

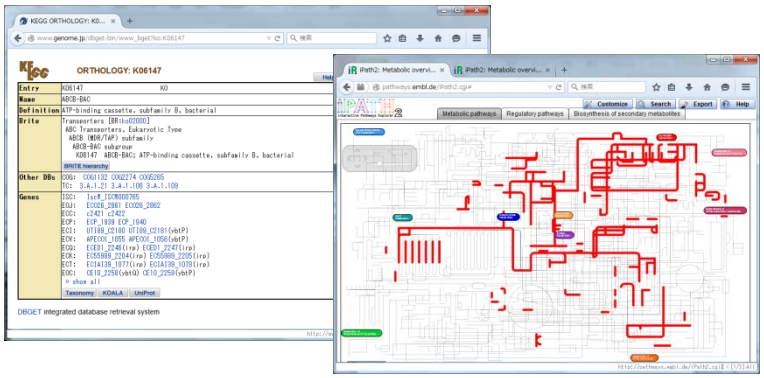


GHOST-MPの利用方法

GHOST-MPのメタゲノムデータに対する相同性検索結果の解析



分類群に基づく解析



遺伝子・機能に基づく解析

GHOST-MPについて

監修・製作：東京工業大学 秋山研究室 GHOST-MPチーム
秋山 泰、石田 貴士、角田 将典、鈴木 脩司

資料製作　：（株）情報数理バイオ

ghost-mp@bi.cs.titech.ac.jp

2015年3月

独立行政法人理化学研究所
HPCI計算生命科学推進プログラム