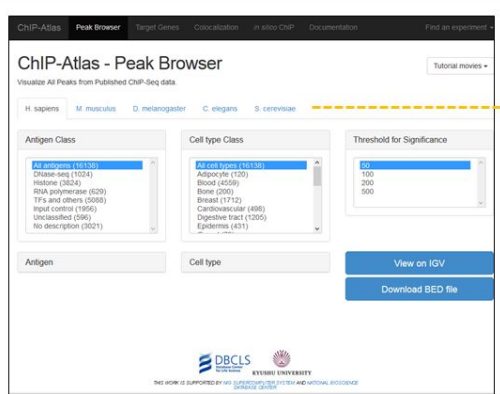


1. チャレンジ課題

チャレンジ: DNA配列からクロマチン特徴を予測



九大沖博士, DBCLS大田博士のDDBJ SRAのクロマチン特徴注釈DB
「ChIP-Atlasデータベース」<http://chip-atlas.org/>

■ DNA配列から特徴予測
ChIP-Atlasピーク領域 = 特徴有DNA領域
↓
予測モデル構築
↓
DNA配列の特徴有無を予測へ

■ 解析対象の生物種

H. sapiens M. musculus D. melanogaster C. elegans S. cerevisiae

ヒト、マウス、ハエ、線虫、酵母
チャレンジは、非公開生物種の「植物」が対象。

2. 訓練・テストのデータファイル

<データ形式>

- 変数(入力: seq_tr, 出力: out_tr)を使って機械学習モデリングを行います。
- 変数(入力: seq_te)を使って出力の予測します。
- 予測データをビッグデータ大学に投稿してください。

変数	行列データ数	データ説明
seq_tr	[(訓練)配列数 x logical] [60,000 x 800]	200bp DNA配列(4 digit encode)を訓練数分 0100001000010001 ... [=CGTT...]
seq_te	[(テスト)配列数 x logical] [10,000 x 800]	200bp DNA配列(4 digit encode)をテスト数分
out_tr	[(訓練)データ数 x 出力条件数] [60,000 x 8]	入力配列が<組織条件><抗体条件>でクロマチン特徴領域を含むか(True)、含まないか(False)

<設置場所>

- *ビッグデータ大学内
- *遺伝研スパコン内 /home/challenge/data/下に設置

<設置ファイル>

DDBJ-challenge.mat	チャレンジの課題データ (MATLABバイナリ形式) .mat読込はR : R.matlab のreadMat(), Python : scipy.io.loadmat()
challenge_ldatest.m	線形判別分析のMATLABスクリプト
challenge_ldatest.sh	challenge_ldatest.mをNIGスパコンにジョブ投入するシェルスクリプト
sample_lda.txt	出力ファイル①ビッグデータ大学に投稿する予測結果ファイル
sample_auc_tr.txt	訓練データによる構築モデル性能評価 : AUC値 AUC(Area under the curve)=受信者動作特性(ROC: Receiver Operating Characteristic)曲線下の面積。モデル構築によるアルゴリズムの性能評価に用いられる。AUC値が1に近い程、性能が高い。

3. 遺伝研スパコンへのログイン

①公開鍵の登録

スパコンwebsiteにて、公開鍵(id_rsa.pub)内容をコピーして、入力フォームにペースト。

登録後、数分でログイン可能になる。

入力フォームURL : <https://sc.ddbj.nig.ac.jp/index.php/ja-form-ssh-application-2>

②sshでスパコンゲートウェイに接続

ssh -i .ssh/id_rsa アカウント名@gw2.ddbj.nig.ac.jp

③計算ノードにログイン

* **qlogin** (通常利用の場合 : ノード名指定不要)

もしくは

* **ssh nt161** (チャレンジ専用GPUノードでmatlabを対話モードで起動する場合 :
nt161~nt170の10台中から1台を指定してsshでログイン)

4. MATLABの線形判別分析法で、Challengeデータを解析

遺伝研スパコンでの実行方法

```
[kaminuma@nt170 160706]$ cp /home/challenge/data/challenge_ldatest.* .
```

```
[kaminuma@nt170 160706]$ ls -l
```

```
合計 16
```

```
drwxr-xr-x 2 kaminuma yn-nig 4096 7月 5 18:45 2016 ./
drwxrwx--- 12 kaminuma yn-nig 4096 7月 5 18:44 2016 ../
-rw-r--r-- 1 kaminuma yn-nig 1346 7月 5 18:46 2016 challenge_ldatest.m
-rw-r--r-- 1 kaminuma yn-nig 122 7月 5 18:46 2016 challenge_ldatest.sh
```

```
[kaminuma@nt170 160706]$ cat challenge_ldatest.sh
```

```
#!/bin/bash
#$ -S /bin/bash
#$ -cwd
#$ -l mem_req=24G
#$ -l s_vmem=24G
#$ -l challenge
```

```
matlab -nodisplay -r "challenge_ldatest"
```

```
[kaminuma@nt170 160706]$ less challenge_ldatest.m
```

```
function challenge_ldatest
%
% DDBJ data analysis challenge
%
% Eli Kaminuma
% 2016/7/6 DDBJing sample code
%
%----- data load-----

tmp=load('/home/challenge/data/DDBJ-challenge.mat');

seq_tr=double(tmp.seq_tr); % training input sequence
seq_te=double(tmp.seq_te); % test input sequence
ans_tr=double(tmp.out_tr); % training answer

%----- modeling/predicting individual eight conditions -----

AUCLOG=zeros(8,1);
TimeLOG=zeros(8,1);
N_TR=size(seq_tr,1)
N_TE=size(seq_te,1)

out=zeros(N_TE,8);
```

```

for kk=1:8

    %-----
    tic; % timer

    out_tr=ans_tr(:,kk); % pickup kk-th training data

    %----modeling/calculating training performance-----

    model = fitcdiscr(seq_tr, out_tr,'discrimType','Linear');

    [label_tr,score_lda_tr]=predict(model,seq_tr);
    [XT,YT,TT,AUCLOG(kk)] = perfcurve(out_tr,score_lda_tr(:,2),1);

    %----predicting test data-----

    [label_te,score_lda_te]=predict(model,seq_te);
    out(:,kk)=score_lda_te(:,2);

    %----- timer -----

    TimeLOG(kk) = toc;

end

dlmwrite('sample_lda.txt',out,' ');
dlmwrite('sample_auc_tr.txt',AUCLOG);

fprintf('---FIN[ElapsedTime(min)=%.2f][AUC=%.2f]---\n',sum(TimeLOG)/60,mean(AUCLOG));

%=====[FIN]=====

```

```

[kaminuma@nt170 160706]$ qsub challenge_ldatest.sh
Your job 7147030 ("challenge_ldatest.sh") has been submitted

```

```

[kaminuma@nt170 160706]$ qstat
job-ID prior name user state submit/start at queue jclass
slots ja-task-ID
-----

```

```
7146816 0.25194 QLOGIN kaminuma r 07/05/2016 17:30:38 login.q@nt099i
1
7147030 0.25000 challenge_ kaminuma r 07/05/2016 18:57:10
challenge.q@nt168i 1
```

```
[kaminuma@nt170 160706]$ ls -l
```

```
-rw-r--r-- 1 kaminuma yn-nig 1346 7月 5 18:46 2016 challenge_ldatest.m
-rw-r--r-- 1 kaminuma yn-nig 111 7月 5 19:00 2016 challenge_ldatest.sh
-rw-r--r-- 1 kaminuma yn-nig 0 7月 5 19:00 2016 challenge_ldatest.sh.e7147033 (標
準エラー)
-rw-r--r-- 1 kaminuma yn-nig 898 7月 5 19:02 2016 challenge_ldatest.sh.o7147033 (標
準出力)
-rw-r--r-- 1 kaminuma yn-nig 64 7月 5 19:02 2016 sample_auc_tr.txt
-rw-r--r-- 1 kaminuma yn-nig 652528 7月 5 19:02 2016 sample_lda.txt
```

```
[kaminuma@nt170 160706]$ cat challenge_ldatest.sh.o7147033
```

```
< M A T L A B (R) >
Copyright 1984-2016 The MathWorks, Inc.
R2016a (9.0.0.341360) 64-bit (glnxa64)
February 11, 2016

< 中略 >
N_TR =
    60000

N_TE =
    10000

---FIN [ElapsedTime(min)=1.36][AUC=0.67]---
    ↑ 総計算時間は1.36分。訓練データ8条件の平均AUCは0.67
```

```
[kaminuma@nt170 160706]$ cat sample_auc_tr.txt
```

```
0.70056
0.77772
0.66623
```

0.64251
0.63793
0.76049
0.58816
0.62413

↑ 訓練データ1条件ずつのAUC, 平均は標準出力での0.67。

```
[kaminuma@nt170 160706]$ cat sample_lda.txt (予測結果 : 真確率ファイル)
```

```
0.044468 0.10076 0.08073 0.48344 0.34521 0.40086 0.30471 0.17754  
0.033496 0.13188 0.20314 0.55447 0.18272 0.13753 0.18789 0.1152  
0.045572 0.428 0.054755 0.30207 0.1341 0.087429 0.18152 0.21076  
0.011437 0.039643 0.17207 0.62627 0.31328 0.4152 0.18731 0.12842  
0.098373 0.31975 0.10752 0.39901 0.12401 0.073645 0.25709 0.21598  
0.048972 0.14751 0.046383 0.4391 0.2288 0.29228 0.28766 0.18423  
0.015213 0.018333 0.23794 0.66641 0.39734 0.52767 0.24068 0.10958  
0.026631 0.18165 0.11912 0.47957 0.16453 0.24346 0.341 0.20033  
0.076155 0.090377 0.2123 0.57961 0.3427 0.26133 0.28921 0.2449  
0.017996 0.10757 0.070729 0.39395 0.35233 0.24187 0.26481 0.12811
```

:

<中略>

5. 予測結果をビッグデータ大学に投稿

NIGスパコンからローカルPCへデータファイルをコピーします。(ローカルPCのターミナルで操作)

最後にピリオドが要るので注意して下さい。

```
[eli@jupitar ~]$ scp kaminuma@gw2.ddbj.nig.ac.jp:~/sample_lda.txt .  
Enter passphrase for key '/home/eli/.ssh/id_rsa':  
sample_lda.txt 100% 637KB 637.2KB/s 00:00
```

ビッグデータ大学のウェブサイトに行きます。

<http://universityofbigdata.net/competition/5749873794088960>



ビッグデータ大学

コンペ一覧

参加案内

[DDBJデータ解析チャレンジ 2016] DNA配列からのクロマチン特徴予測

※このコンペティションはDDBJ Data Analysis Challengeとして開催しております。
詳細は<http://www.ddbj.nig.ac.jp/ddbj-challenge2016-j.html>を御参照ください。
[DDBJデータ解析チャレンジ同意書](#)

DDBJでは、高速DNAシーケンサ由来のビッグデータDDBJ SRAを提供しています。このコンペティションAtlasデータベースを用いて、入力DNA配列に対応するゲノム領域にクロマチン特徴領域が含まれる；オフ機能に関する領域であり、ChIP-Atlasデータベースでのピーク領域に相当します。ChIP-Atlasマーカーが設定されています。課題データは8条件で構成されています。

問題の種類	分類
評価指標	平均ROC-AUC
状態	開催中
開始日時	2016/07/06 00:00 (Japan Standard Time)
終了日時	2016/08/31 23:59 (Japan Standard Time)
公開設定	公開(誰でもコンペティションの内容を閲覧できます)
参加者制限	制限なし(ユーザ登録済みであれば誰でもコンペティションに参加できます)

Googleアカウントでログインします。

予測用データのダウンロード・予測結果の提出

データダウンロード・予測結果提出には、[Googleアカウント](#)によるログインが必要になります。
参加登録がまだの場合は、[参加登録](#)を行ってください。

ローカルPCの[sample_lda.txt](#)を選択して、Submitボタンを押したら投稿完了です。

予測結果の提出

Change Remove Submit



管理者アカウントには提出回数制限はありません

最大20MBまで提出できます。ZIPファイルでの提出も可能です。

追記： 遺伝研スパコンchallenge.qの混雑具合は、[遺伝研スパコンのホームページ](#)から確認可能です。Phase 2研究用thin計算ノードのchallenge.qを御覧ください。
challenge.q以外は研究利用のキューです。









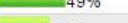

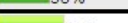



Phase 2

研究用medium計算ノード

qname	host	slot (スロット数)			memory (GB)	
		use / all percent	disabled	in use	req / all percent	
month_medium.q	8	384 / 640  60%	0	2239	16147 / 16384  98%	

研究用thin計算ノード

※month_gpu.qとshort.qは実ノードが同じため、in use(使用メモリ量)が同じ値となります。

qname	host	slot (スロット数)			memory (GB)	
		use / all percent	disabled	in use	req / all percent	
challenge.q	10	0 / 200 0%	0	36	0 / 640 0%	
debug.q	4	1 / 80 1%	0	23	30 / 256  11%	
login_sp.q	2	1 / 60 1%	0	6	12 / 192  6%	
month_gpu.q	62	270 / 310  87%	50	796	520 / 1984  26%	
month_phi.q	30	569 / 600  94%	0	346	791 / 1920  41%	
short.q	62	389 / 930  41%	150	796	1056 / 1984  53%	
week_hdd.q	51	505 / 1020  49%	0	972	3238 / 3264  99%	
week_ssd.q	32	245 / 640  38%	0	643	2035 / 2048  99%	
	265	2034 / 4200  48%	290	3682	7994 / 13440  59%	