FDR 구현 실습

2020.06

김현우

FDR 구현 실습

- **결과 파일 입력 및 charge 별로 구분**
  - **결과 파일의** PSM**을** charge **별로 구분하여 저장하는 코드 작성**

    - 1) **주어진** 201909r.1.real.txt **파일을 입력 받는 코드 작성**

    - 2) **저장될 파일명 코드 작성**
      - charge 2 : 201909r.1.real.charge2.txt
      - charge 3 : 201909r.1.real.charge3.txt
      - charge 4 **이상** : 201909r.1.real.charge4.txt

FDR 구현 실습

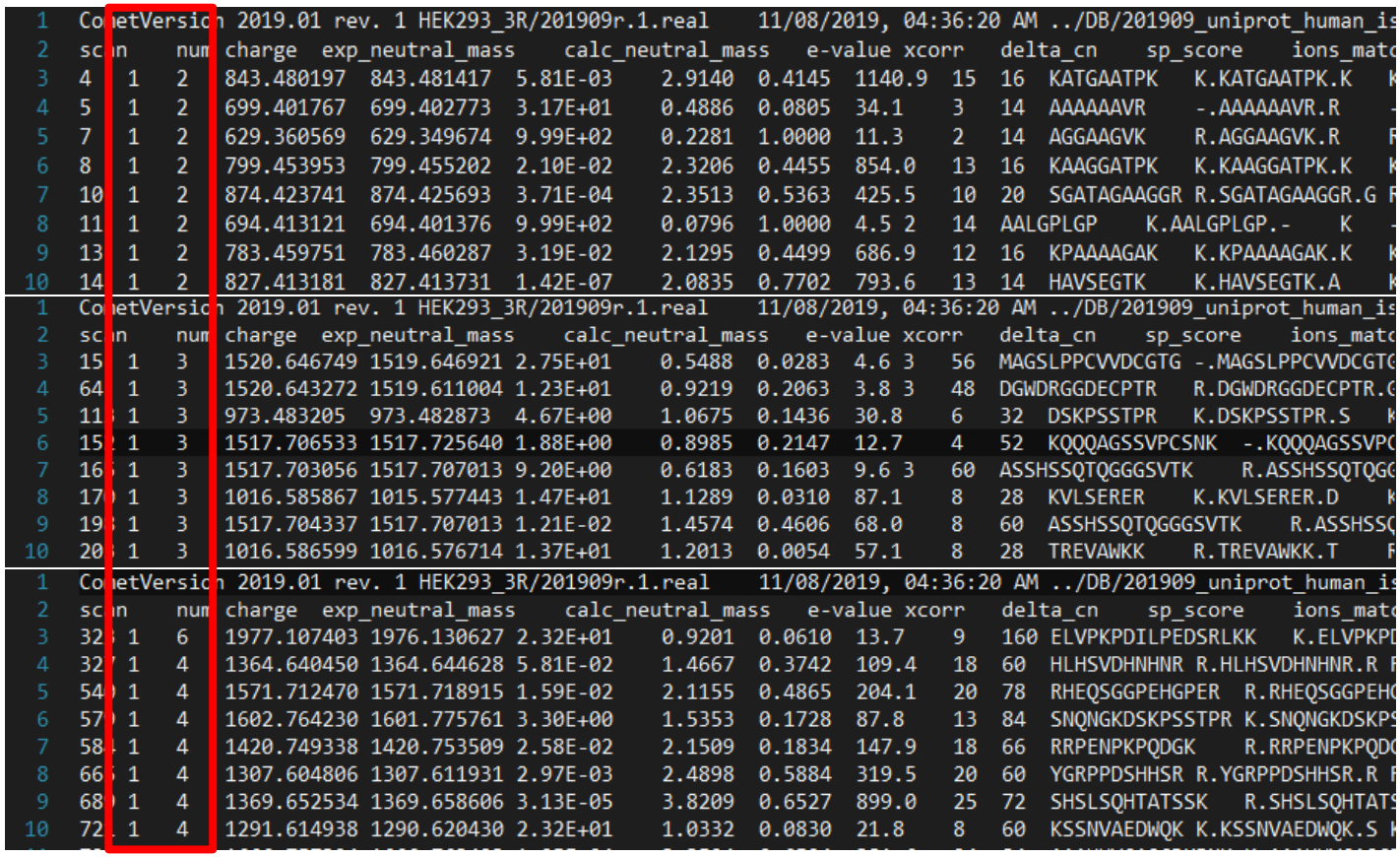
- **결과 파일 입력 및 charge 별로 구분**
  - **결과 파일의** PSM**을** charge **별로 구분하여 저장하는 코드 작성**

    - 3) **조건**
      - 1. charge 2**로 구성된** PSM**으로 구분**
      - 2. charge 3**으로 구성된** PSM**으로 구분**
      - 3. charge 4 **이상으로 구성된** PSM**으로 구분**
      - 4. **각 구분되는** PSM**의** rank**는** 1**등인 것만 저장한다**.

    - 4) **위 네 가지 조건에 맞게 코드를 작성하여 생성된 파일에 저장하는 코드 작성**.

FDR 구현 실습

- **결과 파일 입력 및 charge 별로 구분**
  - 다음과 같은 결과가 나오게 코드 작성



	CometVersion	2019.01 rev. 1	HEK293_3R/201909r.1.real			11/08/2019, 04:36:20 AM	../DB/201909_uniprot_human_is					
	scan	num	charge	exp_neutral_mass	calc_neutral_mass	e-value	xcorr	delta_cn	sp_score	ions_matc		
3	4	1	2	843.480197	843.481417	5.81E-03	2.9140	0.4145	1140.9	15	16	KATGAATPK K.KATGAATPK.K K
4	5	1	2	699.401767	699.402773	3.17E+01	0.4886	0.0805	34.1	3	14	AAAAAAVR -.AAAAAAVR.R -
5	7	1	2	629.360569	629.349674	9.99E+02	0.2281	1.0000	11.3	2	14	AGGAAGVK R.AGGAAGVK.R R
6	8	1	2	799.453953	799.455202	2.10E-02	2.3206	0.4455	854.0	13	16	KAAGGATPK K.KAAGGATPK.K K
7	10	1	2	874.423741	874.425693	3.71E-04	2.3513	0.5363	425.5	10	20	SGATAGAAGGR R.SGATAGAAGGR.G R
8	11	1	2	694.413121	694.401376	9.99E+02	0.0796	1.0000	4.5	2	14	AALGPLGP K.AALGPLGP.- K -
9	13	1	2	783.459751	783.460287	3.19E-02	2.1295	0.4499	686.9	12	16	KPAAAAGAK K.KPAAAAGAK.K K
10	14	1	2	827.413181	827.413731	1.42E-07	2.0835	0.7702	793.6	13	14	HAVSEGTK K.HAVSEGTK.A K
1	CometVersion	2019.01 rev. 1	HEK293_3R/201909r.1.real			11/08/2019, 04:36:20 AM	../DB/201909_uniprot_human_is					
2	scan	num	charge	exp_neutral_mass	calc_neutral_mass	e-value	xcorr	delta_cn	sp_score	ions_matc		
3	15	1	3	1520.646749	1519.646921	2.75E+01	0.5488	0.0283	4.6	3	56	MAGSLPPCVVDCGTG -.MAGSLPPCVVDCGTG
4	64	1	3	1520.643272	1519.611004	1.23E+01	0.9219	0.2063	3.8	3	48	DGWDRGGDECPTR R.DGWDRGGDECPTR.G
5	11	1	3	973.483205	973.482873	4.67E+00	1.0675	0.1436	30.8	6	32	DSKPSSTPR K.DSKPSSTPR.S K
6	15	1	3	1517.706533	1517.725640	1.88E+00	0.8985	0.2147	12.7	4	52	KQQQAGSSVPCSNK -.KQQQAGSSVPC
7	16	1	3	1517.703056	1517.707013	9.20E+00	0.6183	0.1603	9.6	3	60	ASSHSSQTQGGGSVTK R.ASSHSSQTQGG
8	17	1	3	1016.585867	1015.577443	1.47E+01	1.1289	0.0310	87.1	8	28	KVLSERER K.KVLSERER.D K
9	19	1	3	1517.704337	1517.707013	1.21E-02	1.4574	0.4606	68.0	8	60	ASSHSSQTQGGGSVTK R.ASSHSSQ
10	20	1	3	1016.586599	1016.576714	1.37E+01	1.2013	0.0054	57.1	8	28	TREVAWKK R.TREVAWKK.T F
1	CometVersion	2019.01 rev. 1	HEK293_3R/201909r.1.real			11/08/2019, 04:36:20 AM	../DB/201909_uniprot_human_is					
2	scan	num	charge	exp_neutral_mass	calc_neutral_mass	e-value	xcorr	delta_cn	sp_score	ions_matc		
3	32	1	6	1977.107403	1976.130627	2.32E+01	0.9201	0.0610	13.7	9	160	ELVPKPDILPEDSRLKK K.ELVPKPD
4	32	1	4	1364.640450	1364.644628	5.81E-02	1.4667	0.3742	109.4	18	60	HLHSVDHNHNR R.HLHSVDHNHNR.R R
5	54	1	4	1571.712470	1571.718915	1.59E-02	2.1155	0.4865	204.1	20	78	RHEQSGGPEHGPER R.RHEQSGGPEHG
6	57	1	4	1602.764230	1601.775761	3.30E+00	1.5353	0.1728	87.8	13	84	SNQNGKDSKPSSTPR K.SNQNGKDSKPS
7	58	1	4	1420.749338	1420.753509	2.58E-02	2.1509	0.1834	147.9	18	66	RRPENPKPQDGK R.RRPENPKPQDG
8	66	1	4	1307.604806	1307.611931	2.97E-03	2.4898	0.5884	319.5	20	60	YGRPPDSHHSR R.YGRPPDSHHSR.R R
9	68	1	4	1369.652534	1369.658606	3.13E-05	3.8209	0.6527	899.0	25	72	SHSLSQHTATSSK R.SHSLSQHTATS
10	72	1	4	1291.614938	1290.620430	2.32E+01	1.0332	0.0830	21.8	8	60	KSSNVAEDWQK K.KSSNVAEDWQK.S K

FDR 구현 실습

- **FDR**
  - Charge **별로 구분 된 각 파일을** FDR**하는 코드 작성**.

    - 1) charge**별로 구분된 파일 입력 및** FDR **후 입력될 파일 코드 작성**
      - **저장될 파일명**
        - charge 2 : 201909r.1.real.charge2.1%.txt
        - charge 3 : 201909r.1.real.charge3.1%.txt
        - charge 4 **이상** : 201909r.1.real.charge4.1%.txt

FDR 구현 실습

- **FDR**
  - Charge **별로 구분 된 각 파일을** FDR**하는 코드 작성.**
    - 2) spectrum number**를 기준으로 정렬하는 코드 작성**

```
1  CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
2  scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
3  4   1   2   843.480197  843.481417  5.81E-03    2.9140  0.4145  1140.9  15  16  KATGAATPK   K.KATGAATPK.K   K
4  5   1   2   699.401767  699.402773  3.17E+01    0.4886  0.0805  34.1    3   14  AAAAAAVR    -.AAAAAAVR.R    -
5  7   1   2   629.360569  629.349674  9.99E+02    0.2281  1.0000  11.3    2   14  AGGAAGVK    R.AGGAAGVK.R    R
6  8   1   2   799.453953  799.455202  2.10E-02    2.3206  0.4455  854.0   13  16  KAAGGATPK   K.KAAGGATPK.K   K
7  10  1   2   874.423741  874.425693  3.71E-04    2.3513  0.5363  425.5   10  20  SGATAGAAGGR R.SGATAGAAGGR.G R
8  11  1   2   694.413121  694.401376  9.99E+02    0.0796  1.0000  4.5 2   14  AALGPLGP    K.AALGPLGP.-    K   -
9  13  1   2   783.459751  783.460287  3.19E-02    2.1295  0.4499  686.9   12  16  KPAAAAGAK   K.KPAAAAGAK.K   K
10 14  1   2   827.413181  827.413731  1.42E-07    2.0835  0.7702  793.6   13  14  HAVSEGTK    K.HAVSEGTK.A    K
```

```
1  CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
2  scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
3  2   1   2   911.410251  911.424496  1.94E+01    0.2483  0.0052  2.9 1   16  GVGYGMMVV   K.GVGYGMMVV.-   K   -
4  2   1   2   911.408603  910.417831  6.30E+00    0.2470  0.0092  7.1 1   14  KNSSSCTK    K.KNSSSCTK.M    K   M
5  2   1   2   911.409215  911.409708  7.44E+00    0.2483  0.0061  3.7 1   14  HSVPSDDR    R.HSVPSDDR.G    R   G
6  3   1   2   631.353245  630.344923  7.60E+01    0.4143  0.6971  22.1    2   14  KAGGGGGK    K.KAGGGGGK.R    K
7  3   1   2   948.437961  948.440006  6.38E-02    0.6158  0.2802  16.2    2   14  LDSEDKDK    K.LDSEDKDK.E    K
8  4   1   2   843.480197  843.481417  5.81E-03    2.9140  0.4145  1140.9  15  16  KATGAATPK   K.KATGAATPK.K   K
9  5   1   2   699.401767  699.402773  3.17E+01    0.4886  0.0805  34.1    3   14  AAAAAAVR    -.AAAAAAVR.R    -
10 6   1   2   631.354099  630.344923  2.25E+01    0.5418  0.6642  50.5    3   14  KAGGGGGK    K.KAGGGGGK.R    K
```

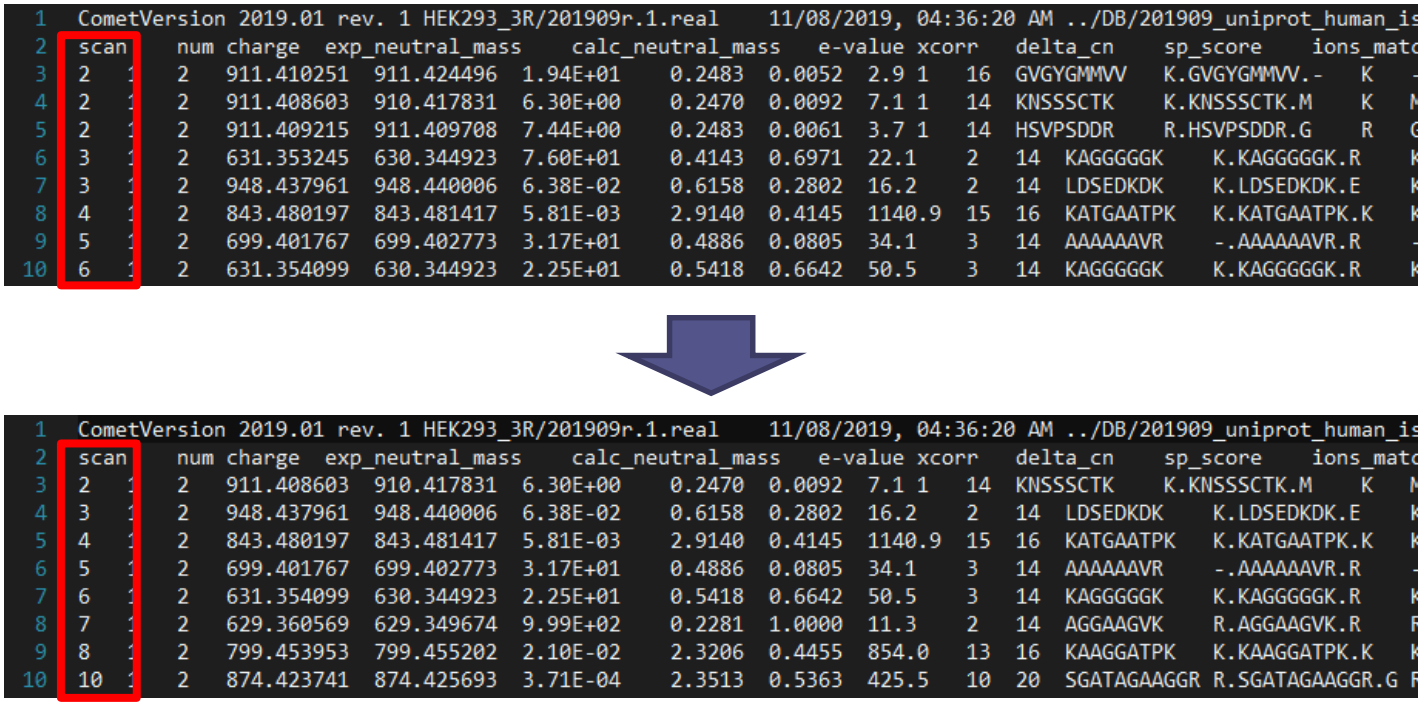
FDR 구현 실습

- **FDR**
  - Charge 별로 구분 된 각 파일을 FDR하는 코드 작성.

    - 3) 중복되는 spectrum number 제거하는 코드 작성
      - 1. spectrum number 가 동일할 때 e-value가 더 낮은 값을 남긴다.

FDR 구현 실습

- **FDR**
  - Charge **별로 구분 된 각 파일을** FDR**하는 코드 작성.**
    - 3) **중복되는** spectrum number **제거하는 코드 작성**



```
CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
2   1   2   911.410251  911.424496  1.94E+01    0.2483  0.0052  2.9 1   16  GVGYGMMVV   K.GVGYGMMVV.-   K   -
2   1   2   911.408603  910.417831  6.30E+00    0.2470  0.0092  7.1 1   14  KNSSSCTK    K.KNSSSCTK.M    K   M
2   1   2   911.409215  911.409708  7.44E+00    0.2483  0.0061  3.7 1   14  HSVPSDDR    R.HSVPSDDR.G    R   G
3   1   2   631.353245  630.344923  7.60E+01    0.4143  0.6971  22.1    2   14  KAGGGGGK    K.KAGGGGGK.R    K
3   1   2   948.437961  948.440006  6.38E-02    0.6158  0.2802  16.2    2   14  LDSEDKDK    K.LDSEDKDK.E    K
4   1   2   843.480197  843.481417  5.81E-03    2.9140  0.4145  1140.9  15  16  KATGAATPK   K.KATGAATPK.K   K
5   1   2   699.401767  699.402773  3.17E+01    0.4886  0.0805  34.1    3   14  AAAAAAVR    -.AAAAAAVR.R    -
6   1   2   631.354099  630.344923  2.25E+01    0.5418  0.6642  50.5    3   14  KAGGGGGK    K.KAGGGGGK.R    K
```

```
CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
2   1   2   911.408603  910.417831  6.30E+00    0.2470  0.0092  7.1 1   14  KNSSSCTK    K.KNSSSCTK.M    K   M
3   1   2   948.437961  948.440006  6.38E-02    0.6158  0.2802  16.2    2   14  LDSEDKDK    K.LDSEDKDK.E    K
4   1   2   843.480197  843.481417  5.81E-03    2.9140  0.4145  1140.9  15  16  KATGAATPK   K.KATGAATPK.K   K
5   1   2   699.401767  699.402773  3.17E+01    0.4886  0.0805  34.1    3   14  AAAAAAVR    -.AAAAAAVR.R    -
6   1   2   631.354099  630.344923  2.25E+01    0.5418  0.6642  50.5    3   14  KAGGGGGK    K.KAGGGGGK.R    K
7   1   2   629.360569  629.349674  9.99E+02    0.2281  1.0000  11.3    2   14  AGGAAGVK    R.AGGAAGVK.R    R
8   1   2   799.453953  799.455202  2.10E-02    2.3206  0.4455  854.0   13  16  KAAGGATPK   K.KAAGGATPK.K   K
10  1   2   874.423741  874.425693  3.71E-04    2.3513  0.5363  425.5   10  20  SGATAGAAGGR R.SGATAGAAGGR.G R
```

FDR 구현 실습

- **FDR**
  - Charge **별로 구분 된 각 파일을** FDR**하는 코드 작성**.
    - 4) e-value **를 기준으로 정렬하는 코드 작성** (e-value**는 낮을 수록 좋음**)

```
1  CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
2  scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
3  2   1   2   911.408603  910.417831  6.30E+00    0.2470  0.0092  7.1 1   14  KNSSSCTK    K.KNSSSCTK.M    K   M
4  3   1   2   948.437961  948.440006  6.38E-02    0.6158  0.2802  16.2    2   14  LDSEDKDK    K.LDSEDKDK.E    K
5  4   1   2   843.480197  843.481417  5.81E-03    2.9140  0.4145  1140.9  15  16  KATGAATPK   K.KATGAATPK.K   K
6  5   1   2   699.401767  699.402773  3.17E+01    0.4886  0.0805  34.1    3   14  AAAAAAVR    -.AAAAAAVR.R    -
7  6   1   2   631.354099  630.344923  2.25E+01    0.5418  0.6642  50.5    3   14  KAGGGGGK    K.KAGGGGGK.R    K
8  7   1   2   629.360569  629.349674  9.99E+02    0.2281  1.0000  11.3    2   14  AGGAAGVK    R.AGGAAGVK.R    R
9  8   1   2   799.453953  799.455202  2.10E-02    2.3206  0.4455  854.0   13  16  KAAGGATPK   K.KAAGGATPK.K   K
10 10  1   2   874.423741  874.425693  3.71E-04    2.3513  0.5363  425.5   10  20  SGATAGAAGGR R.SGATAGAAGGR.G R
```

```
1  CometVersion 2019.01 rev. 1 HEK293_3R/201909r.1.real    11/08/2019, 04:36:20 AM ../DB/201909_uniprot_human_is
2  scan    num charge  exp_neutral_mass    calc_neutral_mass   e-value xcorr   delta_cn    sp_score    ions_matc
3  10713   1   2   2048.769139 2048.777494 6.89E-22    4.6423  0.9381  1300.9  24  36  GGHMDDGGYSMNFNMSSSR R.GGH
4  38257   1   2   1648.883031 1648.889669 9.97E-21    4.0541  0.8518  1087.0  21  28  HDADGQATLLNLLLR R.HDADGQA
5  6296    1   2   1741.658055 1741.663828 1.46E-20    3.6471  0.9315  433.7   17  28  DAHWSEDSEADCHAL R.DAHWSED
6  19151   1   2   2356.138767 2356.148797 1.11E-19    3.1647  0.6911  491.3   17  38  FEAHPNDLYVEGLPENIPFR    K
7  11661   1   2   2254.936863 2254.951553 1.13E-19    4.0485  0.8999  1111.3  23  38  HNDDEQYAWESSAGGSFTVR    K
8  26236   1   2   1648.889865 1648.889669 1.67E-19    4.3308  0.8446  1253.2  22  28  HDADGQATLLNLLLR R.HDADGQA
9  16305   1   2   1754.742039 1754.750475 2.22E-19    2.8261  0.8426  700.2   15  26  NINDAWVCTNDMFR  K.NINDAWV
10 26306   1   2   2040.021579 2040.023596 5.29E-19    4.2786  0.8698  1009.1  22  36  HSSDASSLLPQNILSQTSR K.HSS
```

FDR 구현 실습

- **FDR**
  - Charge 별로 구분 된 각 파일을 FDR하는 코드 작성.

    - 5) FDR 코드 작성
      - PSM의 protein이 target인지 decoy 인지 구분해야한다.
        - Protein에 XXX라는 문자가 포함되면 decoy PSM이라고 판단.
          - ➔ decoy + 1
        - Protein에 XXX라는 문자가 포함되어 있지 않다면, target PSM으로 판단.
          - ➔ target + 1
      - Target과 decoy의 개수를 count하여 FDR 1%(0.01)이내의 target PSM 개수 출력.
        - FDR = number of decoy PSM / number of target PSM

FDR 구현 실습

- **FDR**
  - Charge **별로 구분 된 각 파일을** FDR**하는 코드 작성**.
    - 6) Total target, decoy **개수**, e-value threshold **도 출력하게 코드 작성**.

```
======== RESTART: C:\Users\othertics\Desktop\FDR구현실습\FDR_code.py ========
total tcount = 32905.0
total dcount = 2962.0
e-value threshold = 1.92
total target : 29855.0
>>>
======== RESTART: C:\Users\othertics\Desktop\FDR구현실습\FDR_code.py ========
total tcount = 25539.0
total dcount = 1691.0
e-value threshold = 2.1
total target : 23883.0
>>>
======== RESTART: C:\Users\othertics\Desktop\FDR구현실습\FDR_code.py ========
total tcount = 7457.0
total dcount = 978.0
e-value threshold = 1.04
total target : 6443.0
```

Charge 2

Charge 3

Charge 4 이상