# CHAMPS: Complete Homology Assisted MS Protein Sequencing

Bin Ma

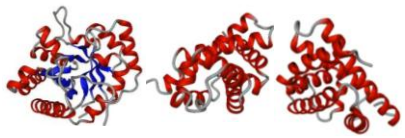
University of Waterloo

# Bottom-Up Proteomics

质谱实验

生物信息

Proteins



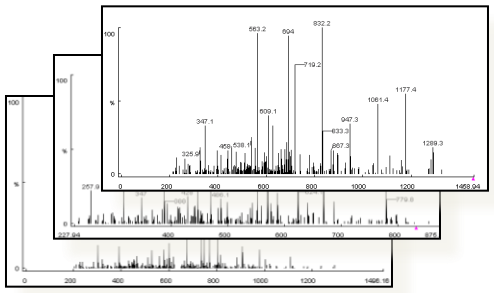
蛋白

digestion

Peptides

多肽

质谱



Peptide ID

Protein ID

```
>protein A                Proteins
PAKGTIRHIHGCDKRGAPWPAS...
>protein B
MSERNHLREIIGNEVR.....
>protein C
LSIMQDKDYSASFIS.....
```

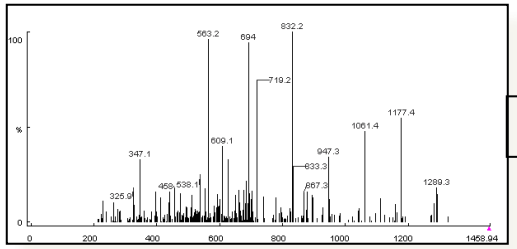
蛋白序列

```
PAK                       Peptides
MSER
LSIMQDK
HIHGCDK
EIIGNEVR
SIMQMDYSASFIS
......
```

肽序列

# 最常见的肽鉴定的两个方式

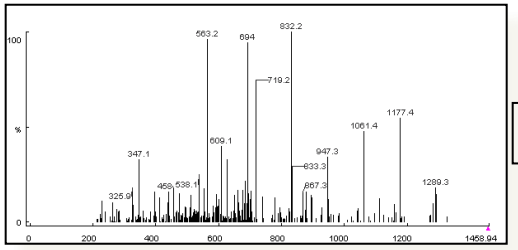
从头测序




*de novo* sequencing

**EISGNEVR**

protein DB



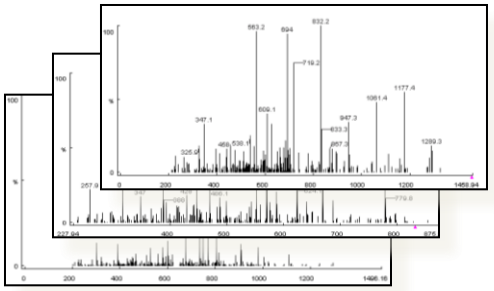

Database search

**ESIGNEVR**

搜库

# 序列库不准确时的问题

MS/MS



Homologous Database

```
>protein A
PAKGTIRHIHGCDKRGDPWRAS…
>protein B
MSERNHLREIIGNEVR……
>protein C
LSIMQDKDYSASFIS……
……
```

>protein A

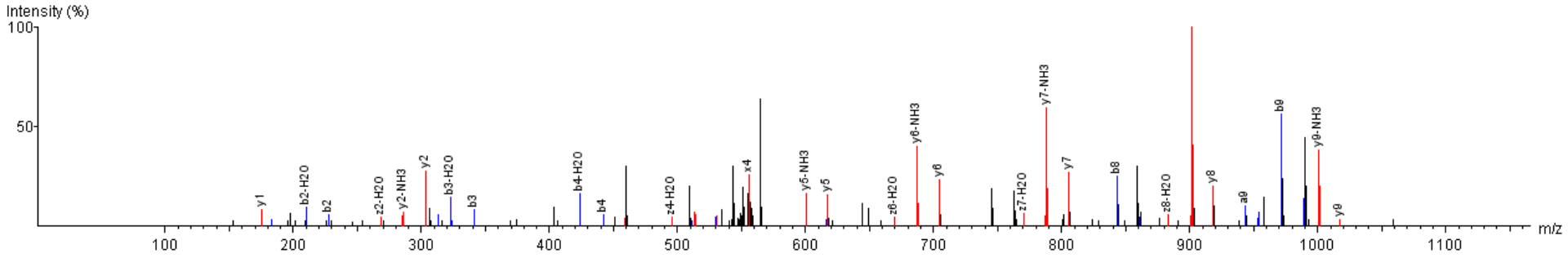
**>protein A**
**PAKGTIRHIHGCDKRGDPWRAS…**

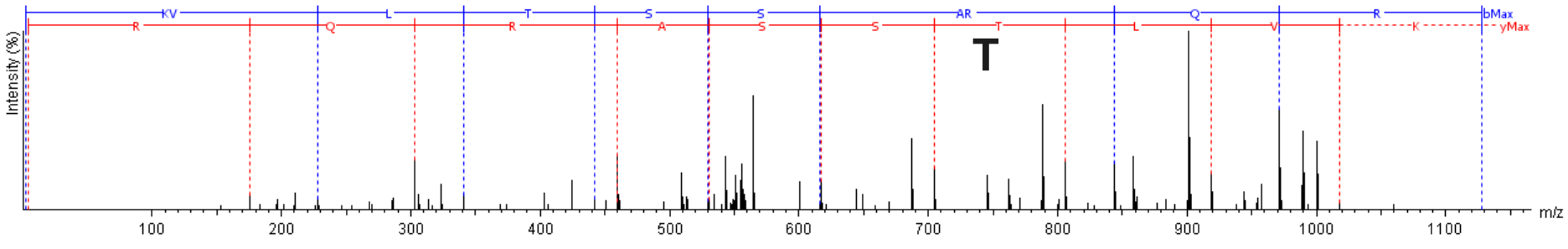
# Example 1: Mutations Between Individuals

214

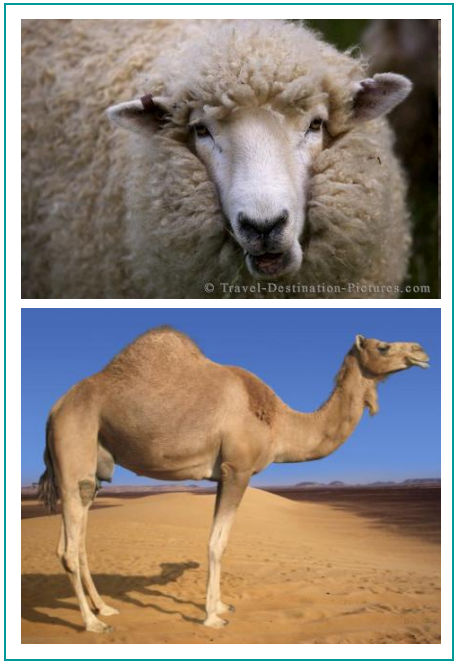
ALBU_BOVIN from swissprot ..MRE KVLASSARQR LRCAS...

Our ALBU_BOVIN ..MRE KVLTSSARQR LRCAS...



# Example 2: Homologous Species

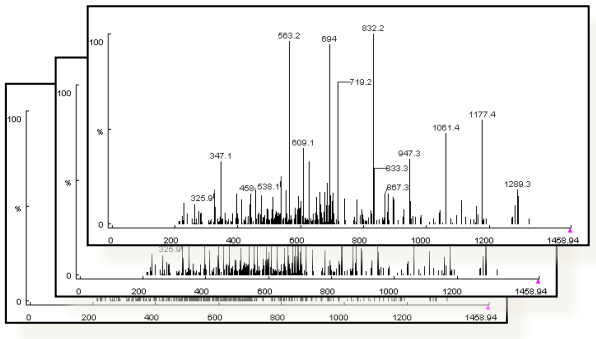

SHEEP DTHKSEIAHRFNDLGEENFQGLVLIAFSQYLQQCPFDEHVKLVKELTEFAK
BOVIN DTHKSEIAHRFKDLGEEHFKGLVLIAFSQYLQQCPFDEHVKLVNELTEFAK

SHEEP KTCVADESHAGCDKSLHTLFGDELCKVATLRETYGDMADCCEKQEPERNEC
BOVIN KTCVADESHAGCEKSLHTLFGDELCKVASLRETYGDMADCCEKQEPERNEC

SHEEP CFLNHKDDSPDLPKLKPEPDTLCAEFKADEKKFWGKYLYEVARRHPYFYAP
BOVIN CFLSHKDDSPDLPKLKPDPNTLCDEFKADEKKFWGKYLYEIARRHPYFYAP

# 我们要挑战的目标

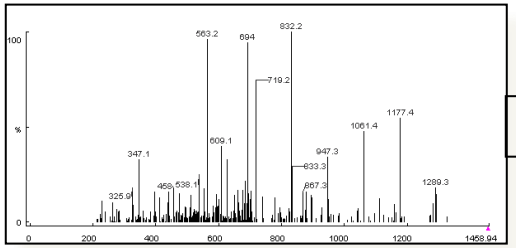
不准确或者同源的
数据库




CHAMPS

全蛋白序列

# 两步走




*de novo* sequencing

EISGNEVR

1

同源库

SPIDER

肽序列

many overlapping de novo sequences

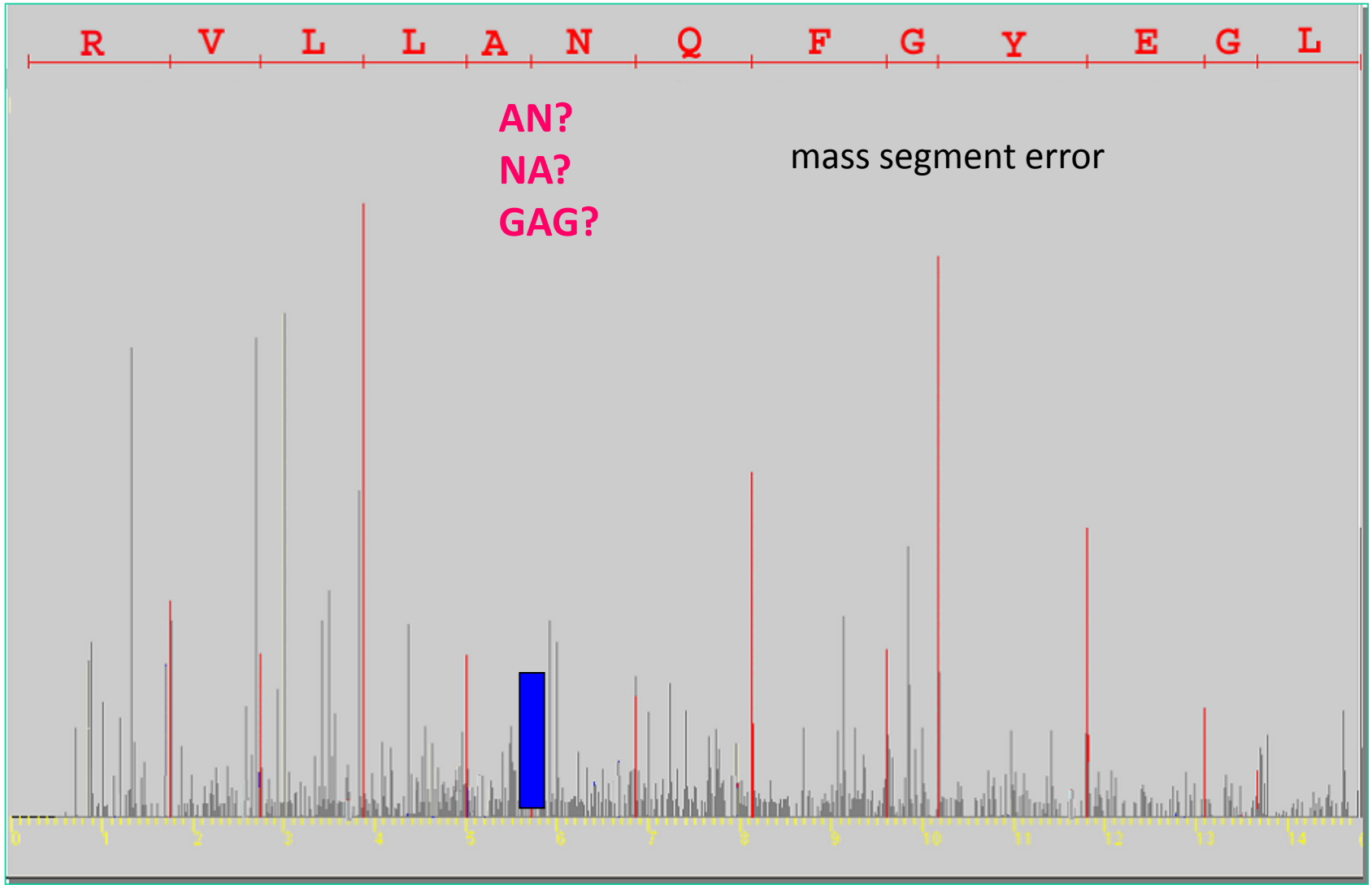
2

同源库

CHAMPS

全蛋白序列

# Common *de novo* sequencing errors



# Mass Segment Error （质量块）

- Most errors are due to incomplete ion ladders in the spectrum.
  - 序列不确定，但质量确定
  - E.g. **HLVLR** v.s. **LHVLR**
- Most de novo sequencing software uses the precursor mass as a constraint.
  - Thus the peptide mass is rarely wrong.

# 传统同源查找的问题

- Suppose the real peptide is **SLCAFK**, and de novo sequencing gives **LSCFAK** with 2 mass segment errors, and homolog is **SLAAFK**.

```
(denovo)   X:    LSCFAK
                       |
(homolog)  Z:    SLAAFK
```

传统查找的解释

```
(denovo)   X:    [LS]C[FA]K
(real)     Y:    [SL]C[AF]K
                  ||    ||  |
(homolog)  Z:    [SL]A[AF]K
```

引入测序误差的解释

# 兼听则明

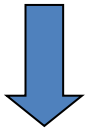
*de novo*

*homolog*

LSCFAK

SLAAFK

bioinformatician



**SLCAFK**

# SPIDER Model

```
(de novo) X:    [LS]C[FA]K
(real)    Y:    [SL]C[AF]K
                 ||   || |
(homolog) Z:    [SL]A[AF]K
```

- Given a de novo  sequence X, and a database sequence Z (two lies).  Try to reconstruct the real sequence Y (the truth).
- The real Y should minimize the de novo errors and the homology mutations needed in the above explanation.

# Two exercises

```
(denovo)   X:   LSCFAV
(real)     Y:   SLCFAV
(homolog)  Z:   SLCF-V
```

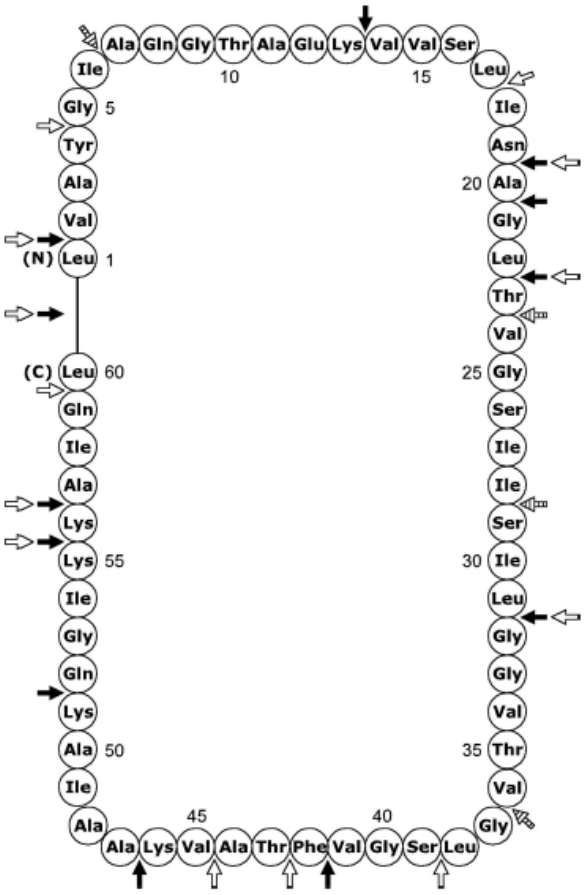
```
(denovo)   X:   LSCFV
(real)     Y:   EACFV
(homolog)  Z:   DACFV
```

*m*(LS)=*m*(EA)=200.1 Da

blosum62

| | C | S | T | P | A | G | N | D | E | Q | H |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C | **9** | -1 | -1 | -3 | 0 | -3 | -3 | -3 | -4 | -3 | -3 |
| S | -1 | **4** | 1 | -1 | 1 | 0 | 1 | 0 | 0 | 0 | -1 |
| T | -1 | 1 | **4** | 1 | -1 | 1 | 0 | 1 | 0 | 0 | -2 |
| P | -3 | -1 | 1 | **7** | -1 | -2 | -1 | -1 | -1 | -1 | -2 |
| A | 0 | 1 | -1 | -1 | **4** | 0 | -1 | -2 | -1 | -1 | -2 |
| G | -3 | 0 | 1 | -2 | 0 | **6** | -2 | -1 | -2 | -2 | -2 |
| N | -3 | 1 | 0 | -2 | -2 | 0 | **6** | 1 | 0 | 0 | -1 |
| D | -3 | 0 | 1 | -1 | -2 | -1 | 1 | **6** | 2 | 0 | -1 |
| E | -4 | 0 | 0 | -1 | -1 | -2 | 0 | 2 | **5** | 2 | 0 |
| Q | -3 | 0 | 0 | -1 | -1 | -2 | 0 | 0 | 2 | **5** | 0 |
| H | -3 | -1 | 0 | -2 | -2 | -2 | 1 | 1 | 0 | 0 | **8** |

# 全蛋白测序




```
Peptide fragment  (1) -AIQLLVAYGIAQGT
                  (2) KAIQLLVAYGIAQGTAEK
                  (3) ------VAYGIAQGTAEKVVSL
                  (4) ---------GIAQGTAEKVVSLINAGL
                  (5) ----------------------INAGLTVGSILSIL
                  (6) ----------------VVSLINAGLTVGSILSILGGVTVGLSGVFTAVK
                  (7) ----------------------------SILSILGGVTVGLSGVFTAVK
                  (8) ---------------------------------------------TAVKAAIAKQGIK
                  (9) ---------------------------------------------VKAAIAKQGIK

Combined sequence     KAIQLLVAYGIAQGTAEKVVSLINAGLTVGSILSILGGVTVGLSGVFTAVKAAIAKQGIK
```

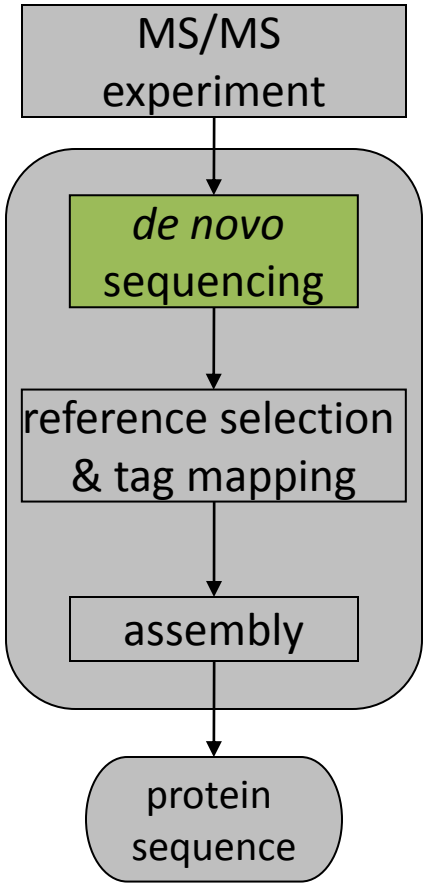
多酶切，形成重叠肽序列
分别测序后拼接

(Figure from Martin-Visscher et al. 2008)

# Existing Works

- Hopper et al., JBC 1989, 106AA
  - trypsin and thermolysin
  - manual de novo sequencing
- Martin-Visscher et al., Appl. Env. Microbiology, 2008, 60AA circular
  - 9 enzymes were used
  - assisted with PEAKS auto de novo sequencing
- Banderia et al. Mol. Cell Proteomics, 2007
  - automated software tool
  - 96% coverage, 90% accuracy
- Banderia et al. Nat. Biotech, 2008
  - improved coverage to 97-99% with reference sequences
  - accuracy not discussed
- Liu et al. Bioinformatics Bioinformatics 2009. (Software name: Champs)
  - achieve >99% accuracy and coverage simultaneously
  - by using reference sequence in a different way

# Champs' Approach (Step 1)




DTHKEELHAR

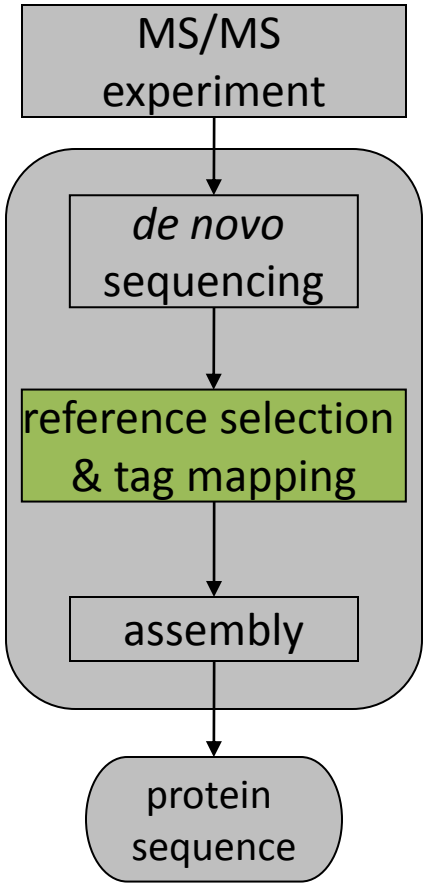
IARHFDDLEFENFQ

QFNGLVLIA

VEPAASQYLQ

**Step 1**. 从头测序产生肽序列，可能有错误。

# Champs' Approach (Step 2)




DTHKSEIAHRFNDLGEENFQGLVLIAFSQYLQ

reference

DTHKEELHAR

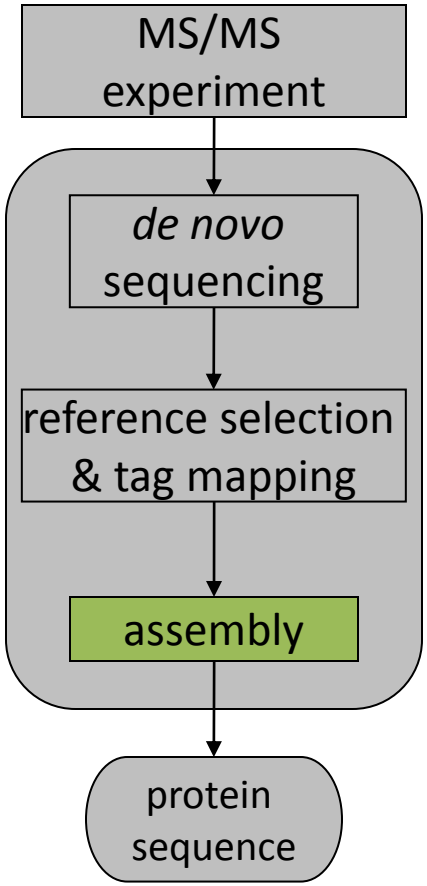
QFNGLVLIA

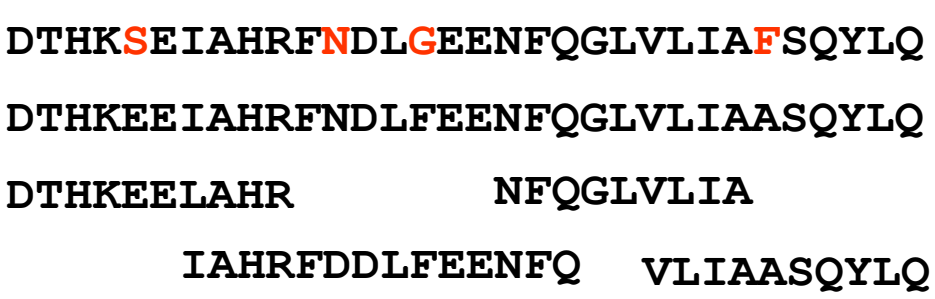
IARHFDDLEFENFQ

VEPAASQYLQ

**Step 2**. SPIDER同源搜索，得到同源蛋白序列，和拼接位置.

# Champs' Approach (Step 3)




**Step 3**. 根据最小化同源错误和测序错误的原则重构蛋白序列。

# Reported Performance

- An experiment was reported in Bioinformatics 2009 25(17): 2174-2180.
- ALBU_BOVIN and LYS_CHICK were used as testing proteins.
  - signal sequences of both proteins removed
  - 583 AA and 129 AA, respectively
- ALBU_BOVIN and LYS_CHICK were removed from swissprot database to form homologous database.
  - Conventional methods can not find the exact protein.
- CHAMPS' performance is the following

| Target Protein | Reference Protein | Reference similarity | Coverage | Accuracy |
|---|---|---|---|---|
| ALBU_BOVIN | ALBU_SHEEP | 92.5% | 99.6% | 100% |
| LYS_CHICK | LYS_COTJA | 95.3% | 100% | 100% |