De novo sequencing in the identification of mass data

Wang Quanhui
Liu Siqi

Beijing Institute of Genomics, CAS

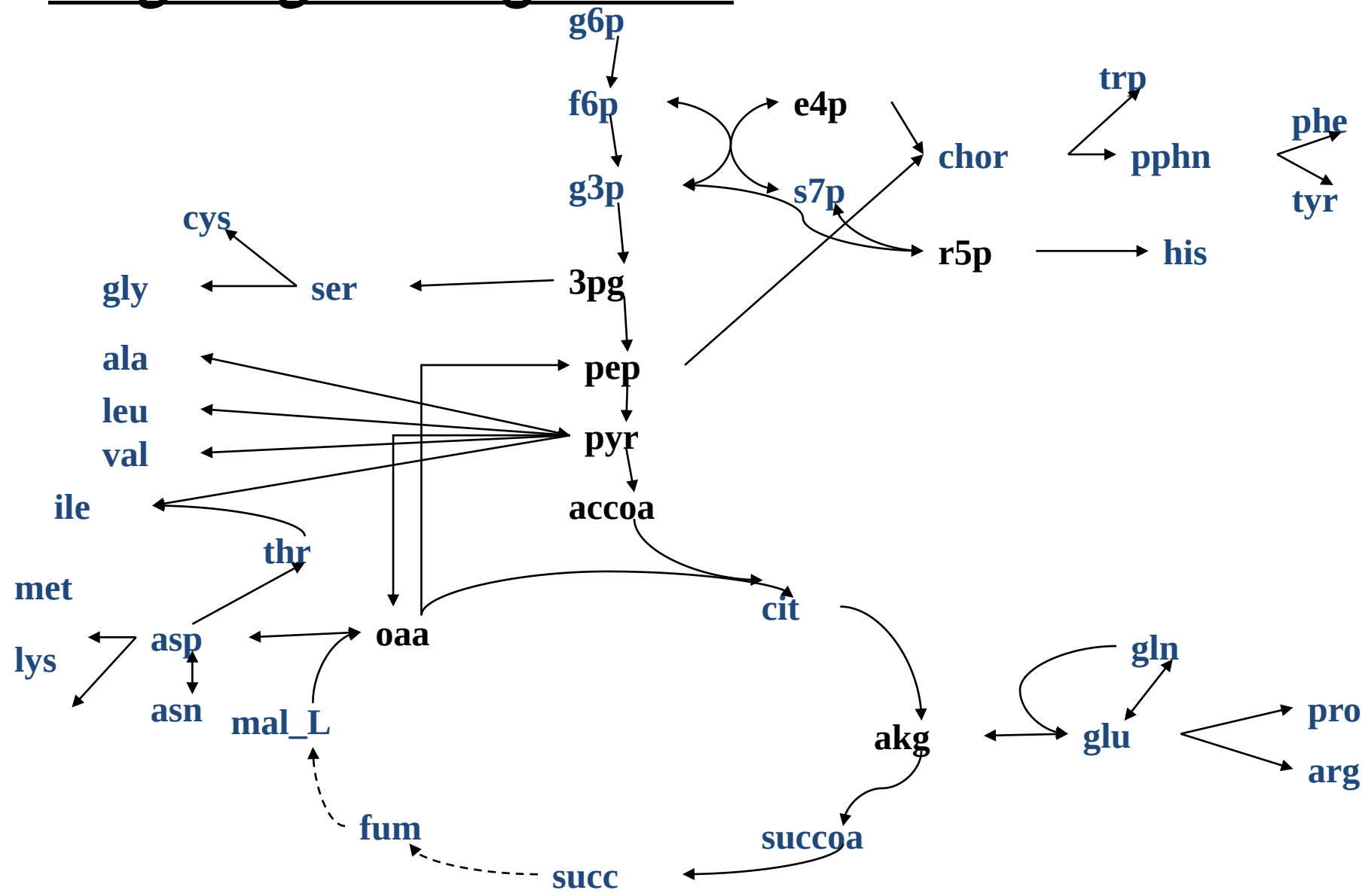
The difficulties in mass data analysis

- Although the techniques of genomic sequencing are being expedited dramatically, there are still a large number of unsequenced genomes, which takes great difficulties for proteome study of those species.

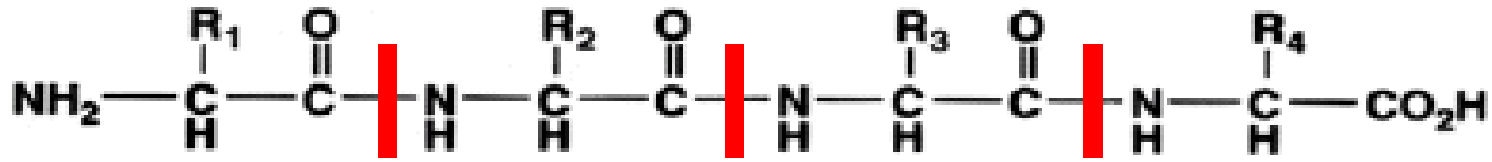
- There are large number of SNPs in genome and unknown PTMs in protein.

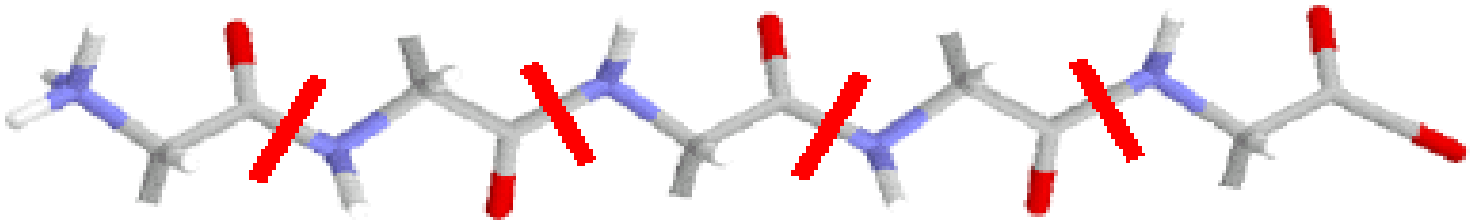
- There are still big percentage of miss-annotation of the sequenced genome.

Large number of miss annotation of *T. tengcongensis* genome



The technique of *de novo* for peptide improved greatly





Fragmentation

Questions

- Is it possible to determine a bacterial proteome by *de novo* with its genomic data being unknown?
- Is it possible to improve protein identification and further correct genome annotation using *de novo*?

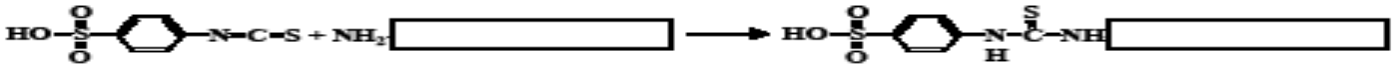
- *De novo* for mass data from the proteins with unknown genes

- *De novo* for mass data from the proteins with predicted genes

- Chemical labeling *de novo*

- Label free *de novo*

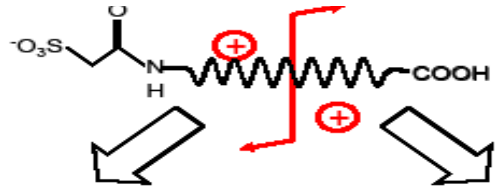
SPITC: A N-terminal sulfonation reagent



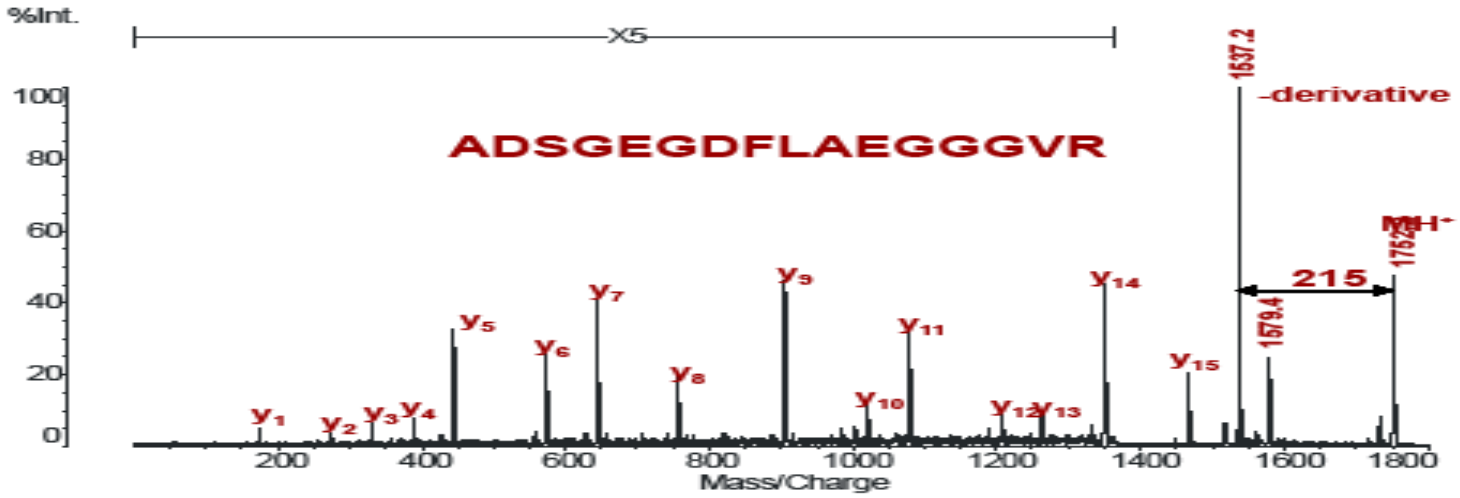
SPITC
4-sulfophenyl isothiocyanate

Sulfonation with SPITC adds **215 Da** to mass of peptide.

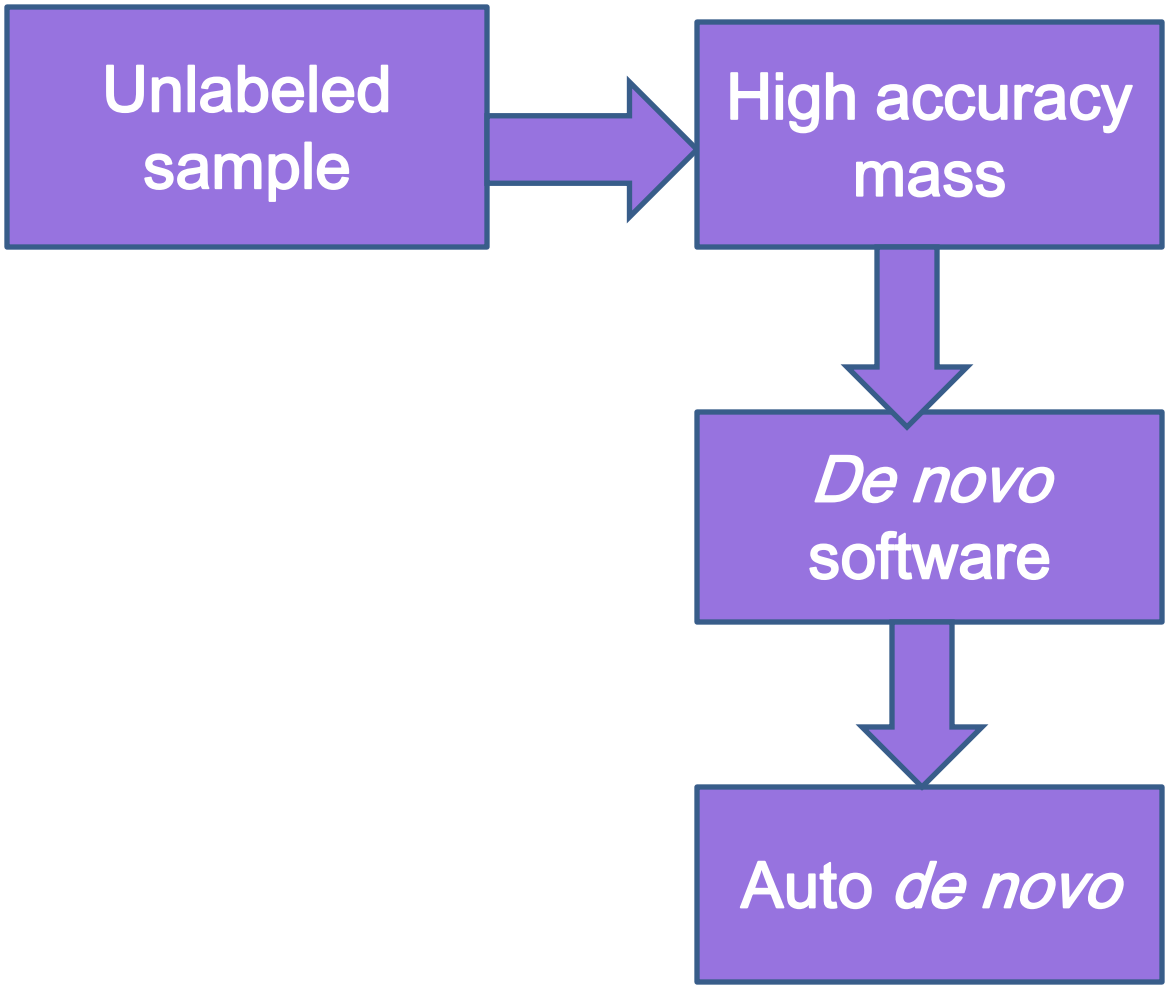
Wang, D.; Kalb, S.R. and Cotter, R.J., *Improved Procedures for N-Terminal Sulfonation of Peptides for Matrix-Assisted Laser Desorption/Ionization Post-Source Decay Peptide Sequencing*, **Rapid Commun. Mass Spectrom. 18** (2004) 96.



- Increased fragmentation
- Easily interpreted PSD spectra!



Label free *de novo*



From an unknown genome to a measurable proteome:

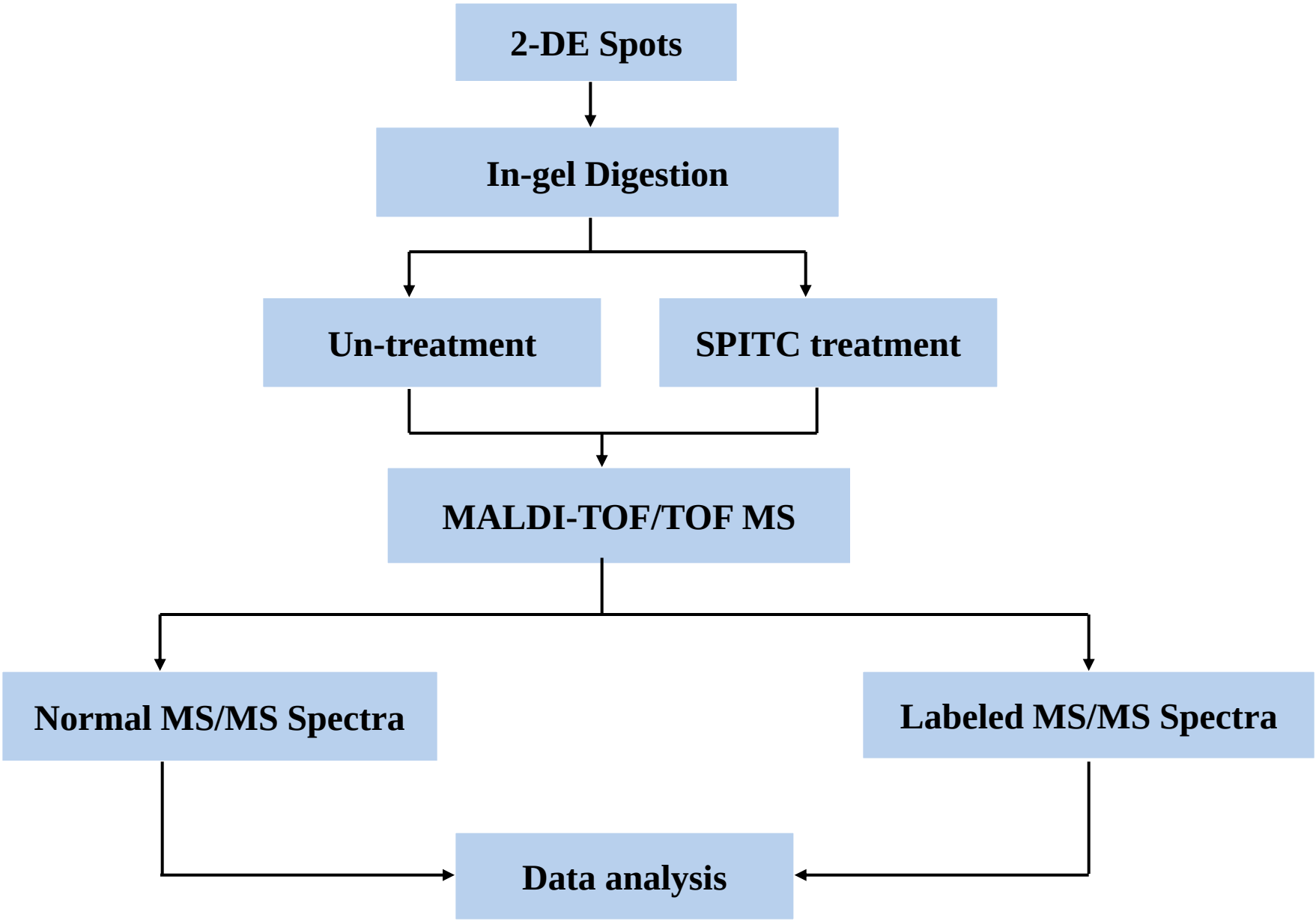
Studying on the pH-dependent proteomes in N10 bacteria

Alkalimonas amylolytica N10 was selected as the target

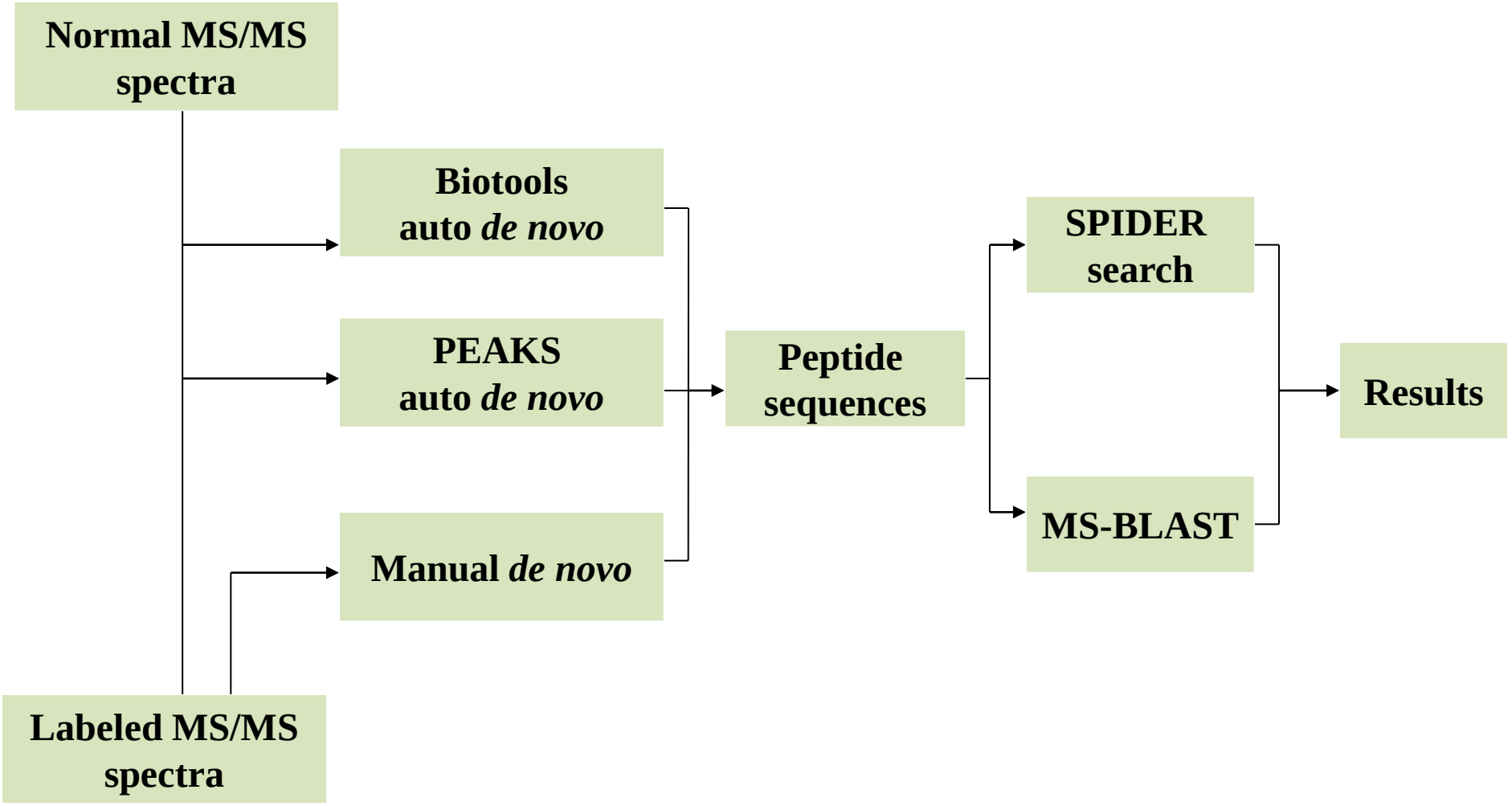
- The N10 bacteria is a kind of gram negative alkaliphilic bacterium, which survives from pH 8.0-11.0 with an optimal pH value is 9.4.

- It is generally accepted that the N10 proteins, especially on membrane, widely respond to the alternations of environmental pH and form the adaptive networks to maintain stable pH in cytoplasm.

The mass spectrometry strategy



The data analysis strategy



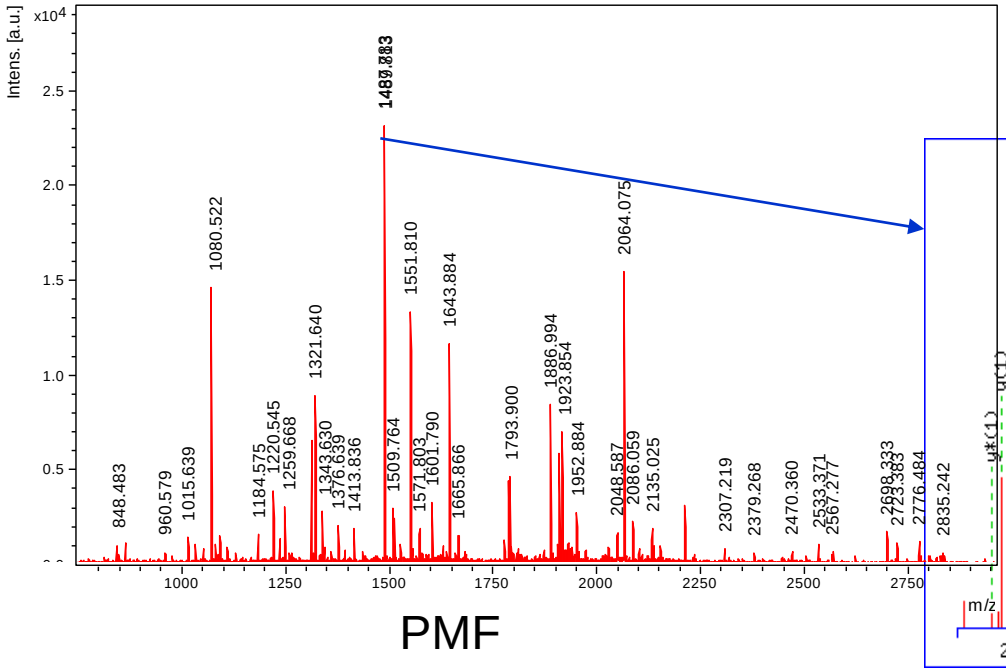
The stringent criteria for *de novo* sequencing

- A deduced sequence should be longer than 7 amino acids.
- A protein should be identified upon at least two unique peptides.
- For MS BLAST, the threshold scores should be higher than 68, 102, 143 and 177 corresponding to high scoring pair values (HSP), 1, 2, 3 and 4, respectively
- All the deduced peptides should be gained from multiple preparation of samples, at least two.

Result 1--identification of differential protein by MALDI TOF/TOF MS

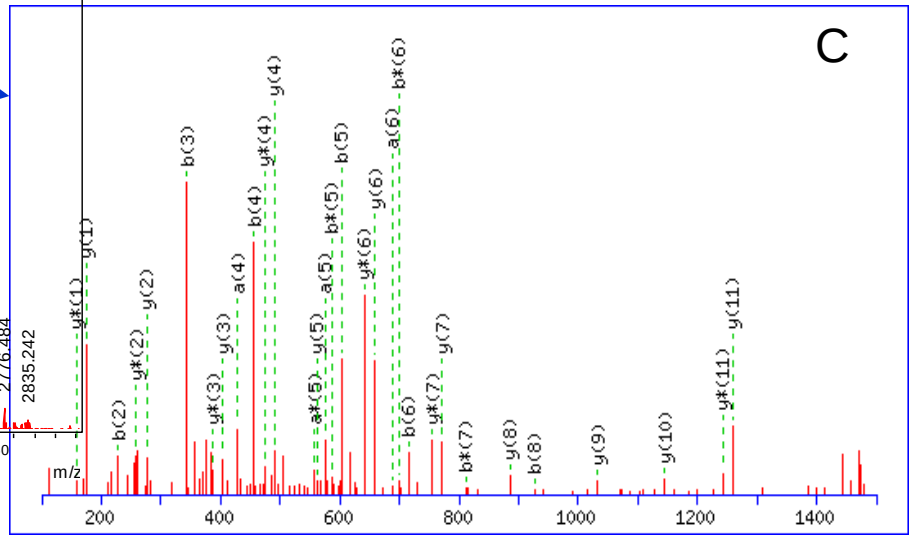
A

B



MS/MS

C

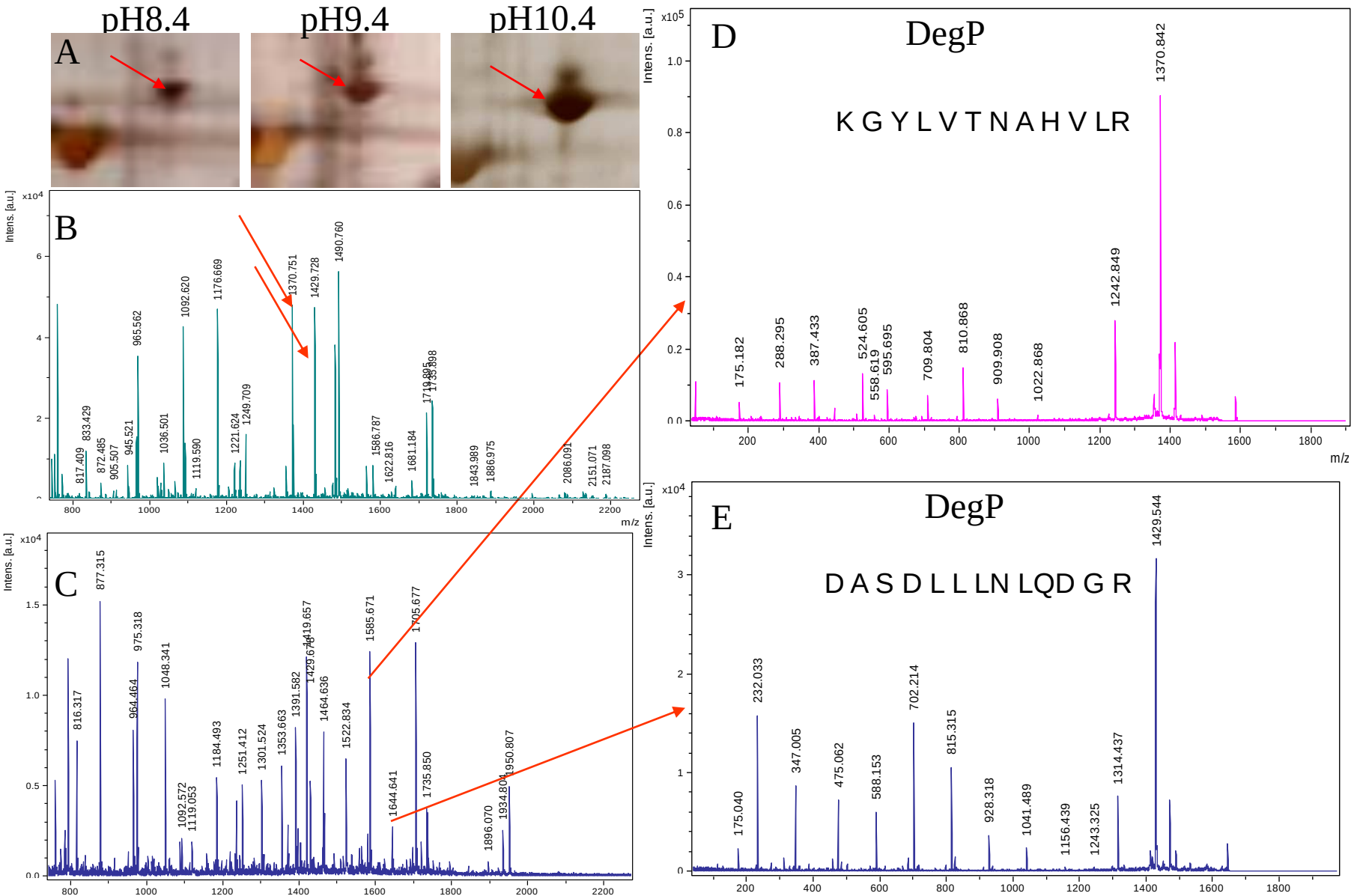


Low identification rates achieved from conventional database-search strategy

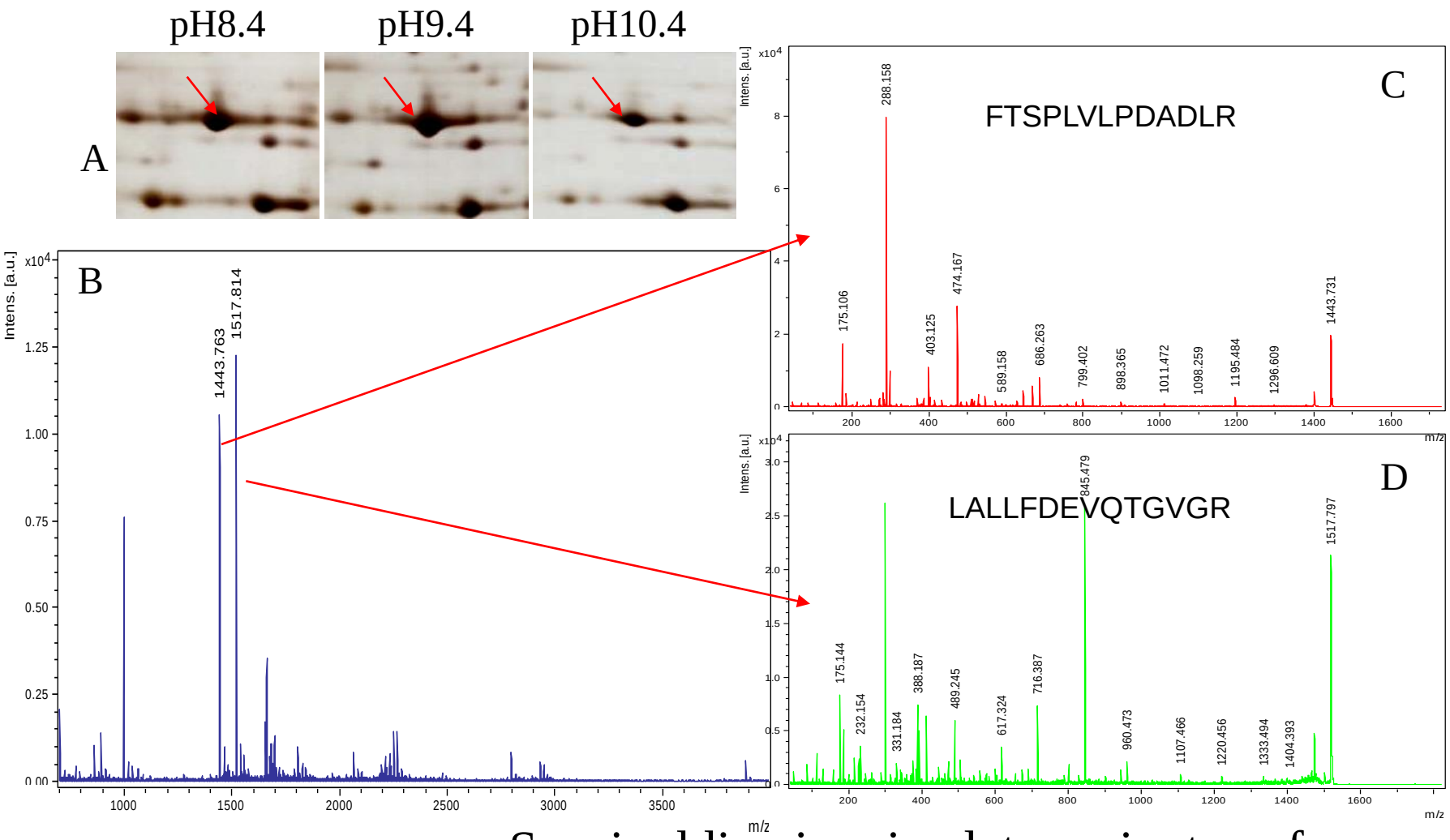
- Statistically, 7 of 26 spots in the membrane fraction and 6 of 46 spots in the cytoplasm were identified as bacterial proteins, respectively.

- Only 13 proteins were identified at the identification rate of 18.1%.

Result 2--Identification of differential proteins by SPITC derivatized *de novo*



Result 3--Identification of differential proteins by underivatized *de novo*

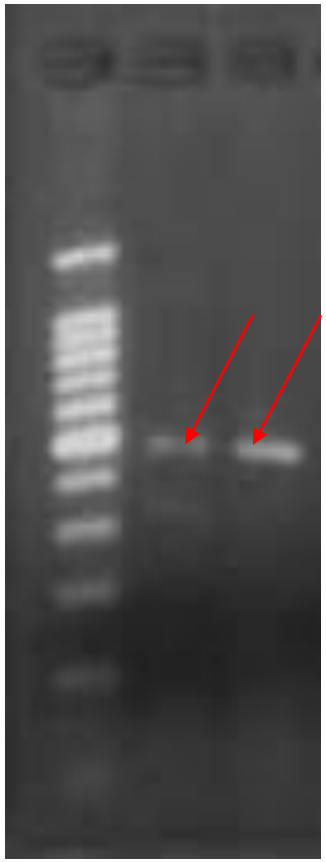


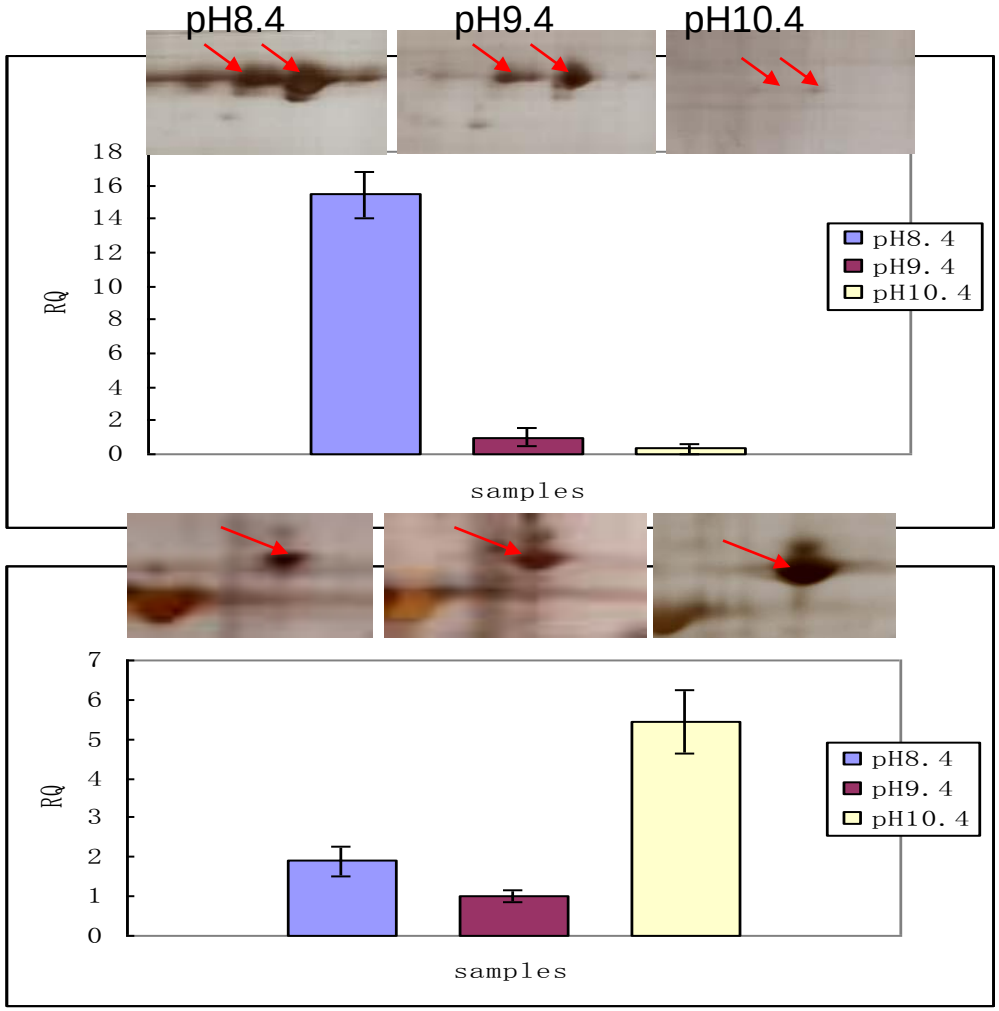
Succinyldiaminopimelate-aminotransferase

Conclusion Table

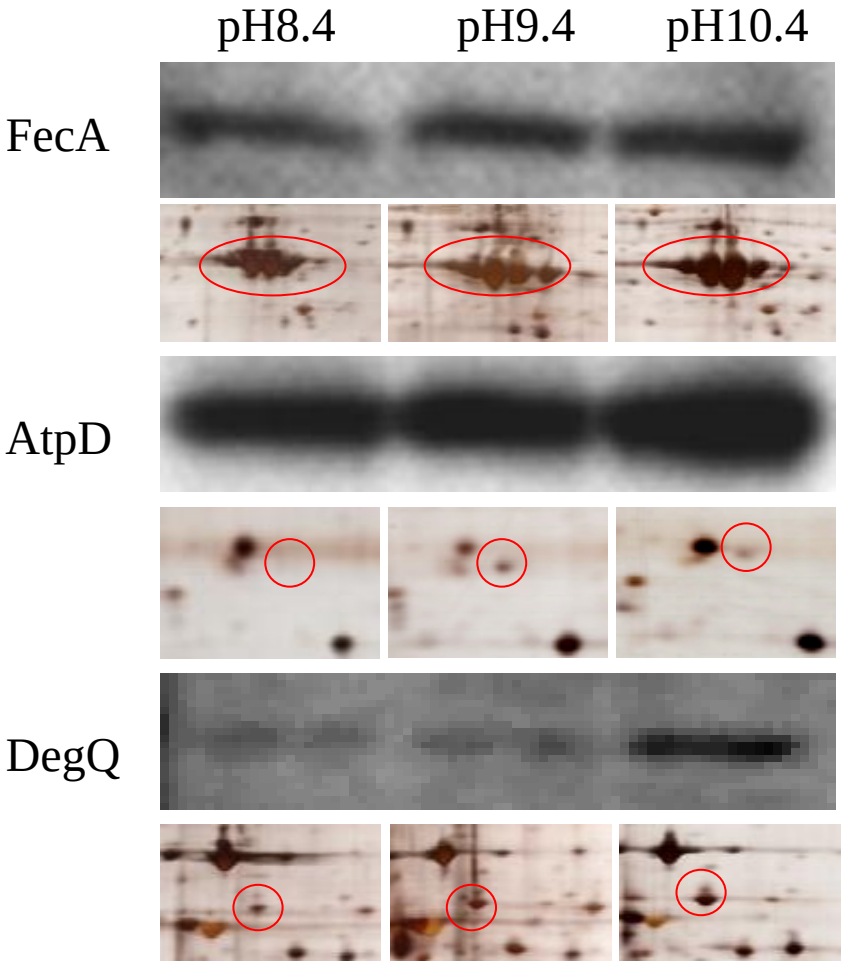
	Differential spots	Mascot search	Normal de novo	SPITC-de novo
Membrane	26	7/72	9/17	10/17
Cytoplasm	46	6/72	13/32	23/32
Total	53	13/72	22/49	33/49
Identification rate	73.6%	18.1%	44.9%	67.3%

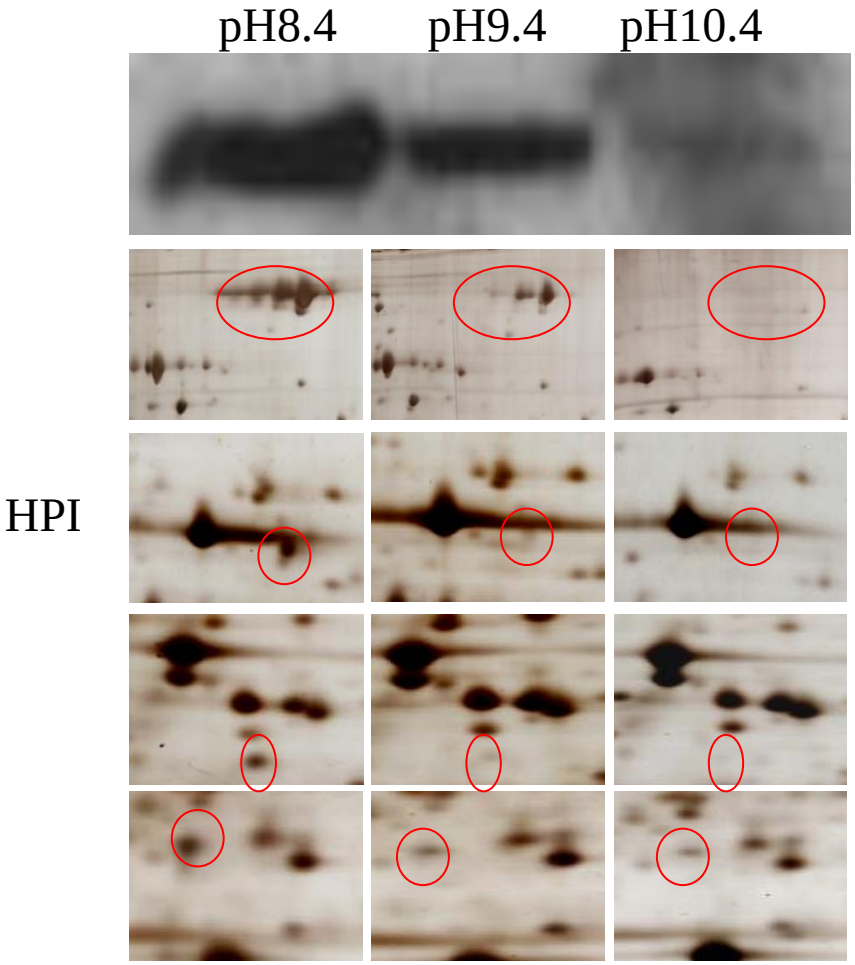
Result 4--Genes of the identified proteins could be amplified and validated by real time PCR



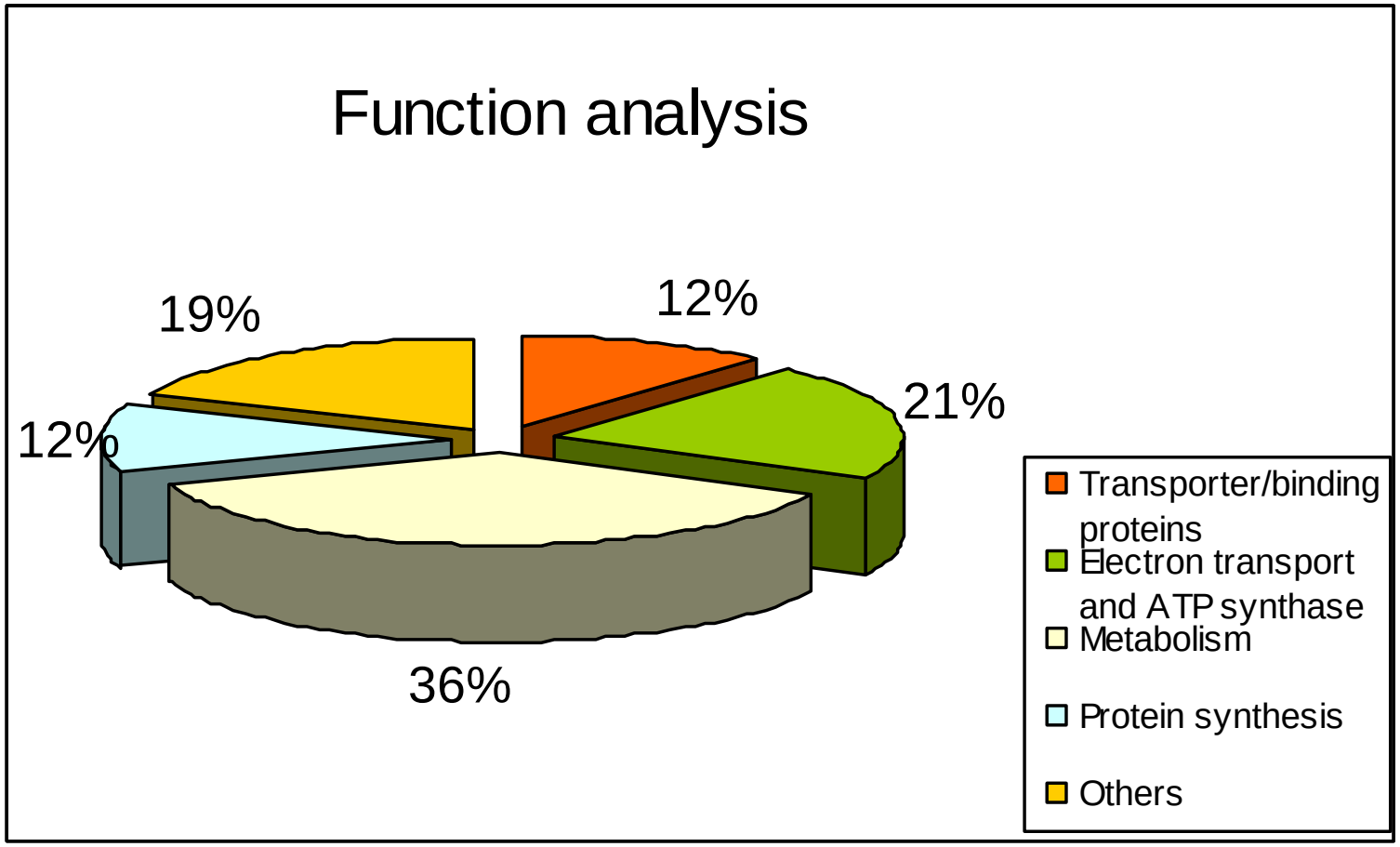


Result 5--Validation of differential proteins by Western blot





Functional analysis of proteins identified



Summaries

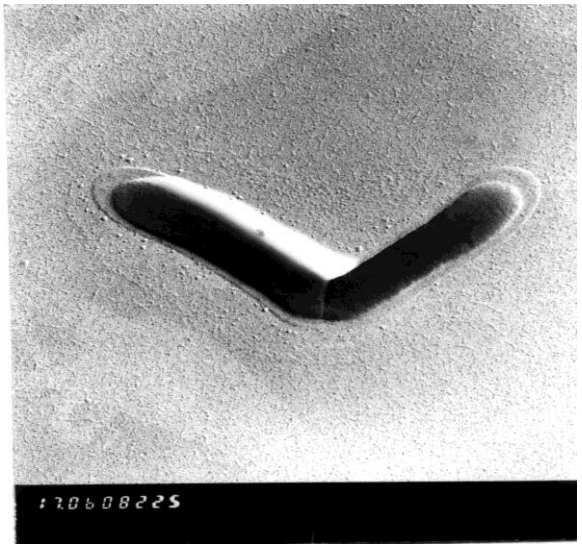
- Based upon the current techniques, a combined strategy for *de novo* sequencing derived from MS/MS signals is feasible and able to achieve accurate identifications;

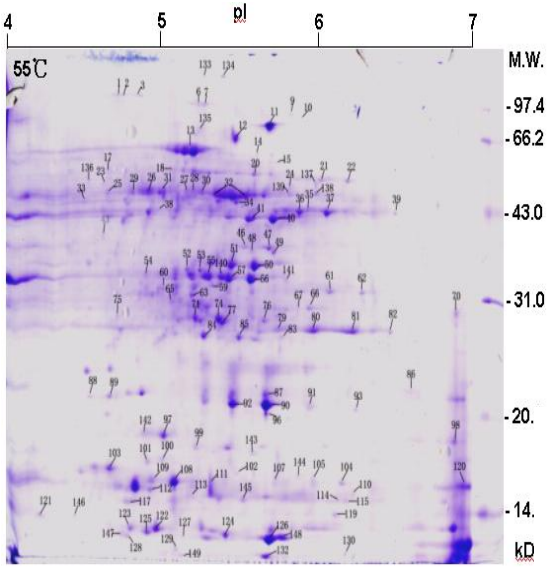
- The *de novo* sequencing is not only successfully in annotation of single proteins, but also useful in proteomic investigation for live species whose genomic data is unavailable, at least for differential proteomics;

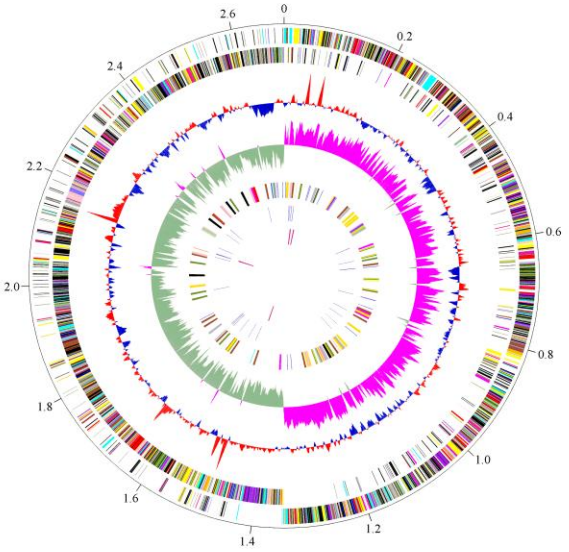
- In the N10 bacteria, membrane and metabolic proteins play the key roles in pH homeostasis within the cells.

De novo sequencing applied in the mass data from species with known genome

Why we choose *TTE*

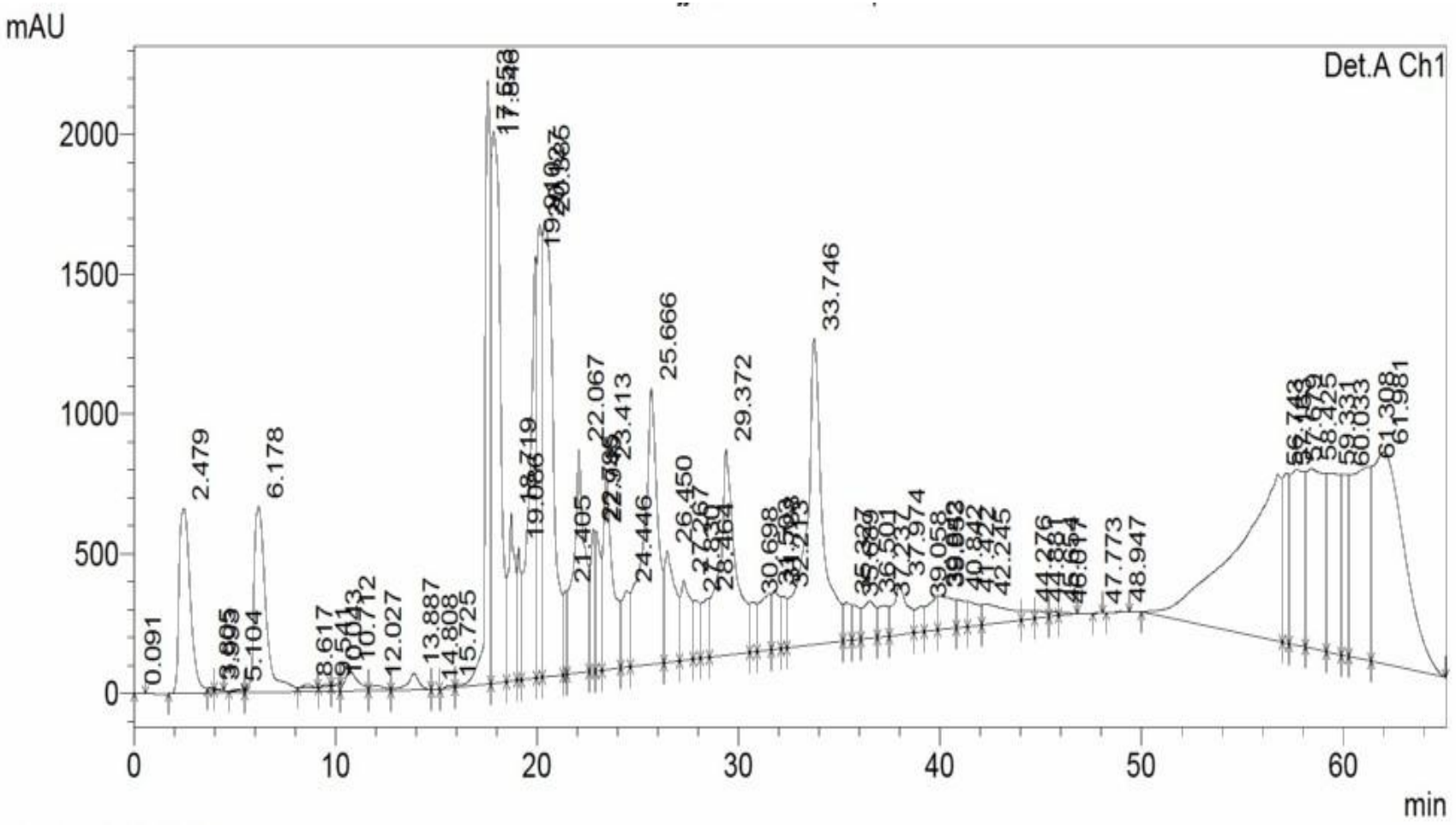




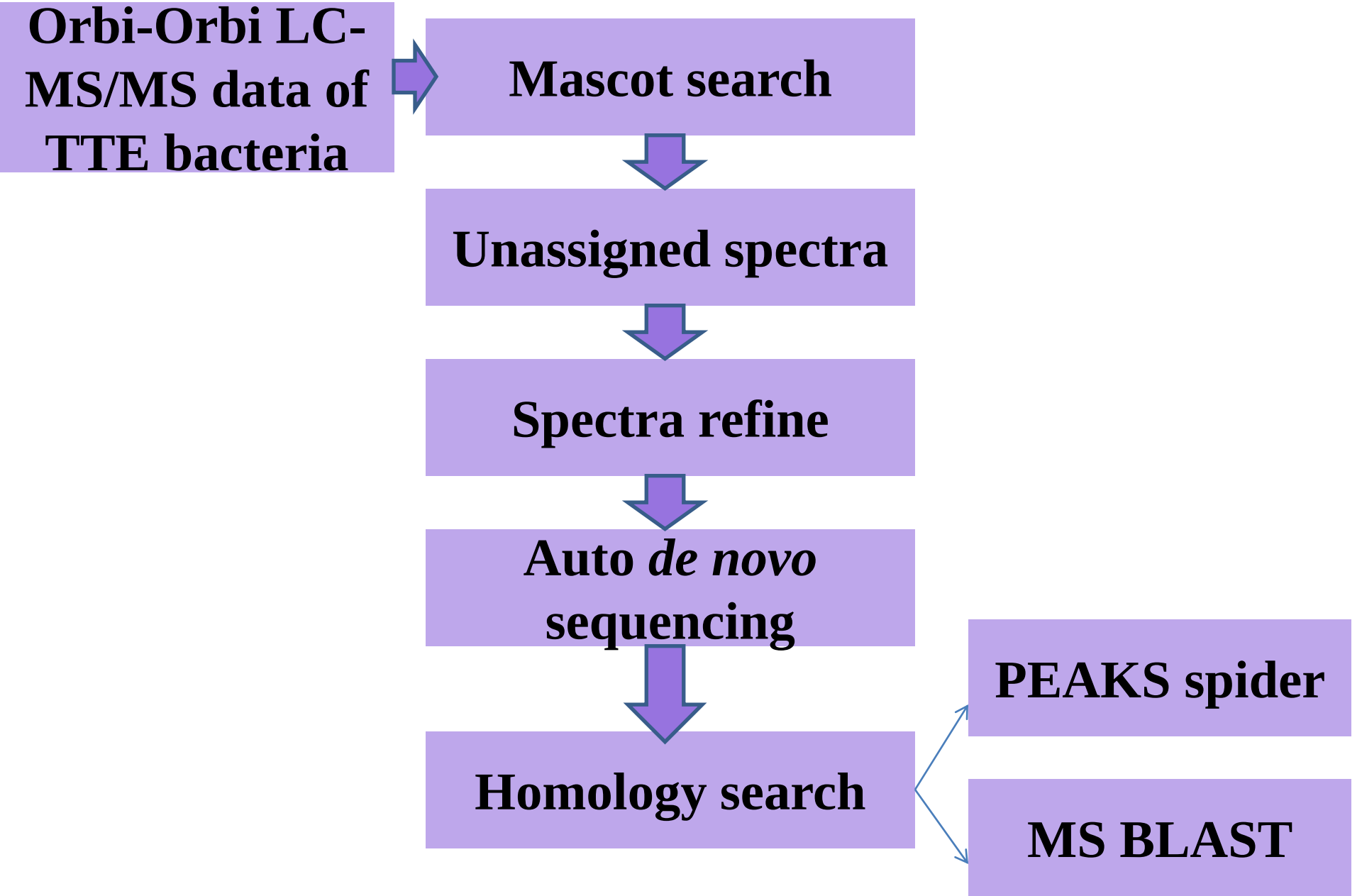


Sample preparation

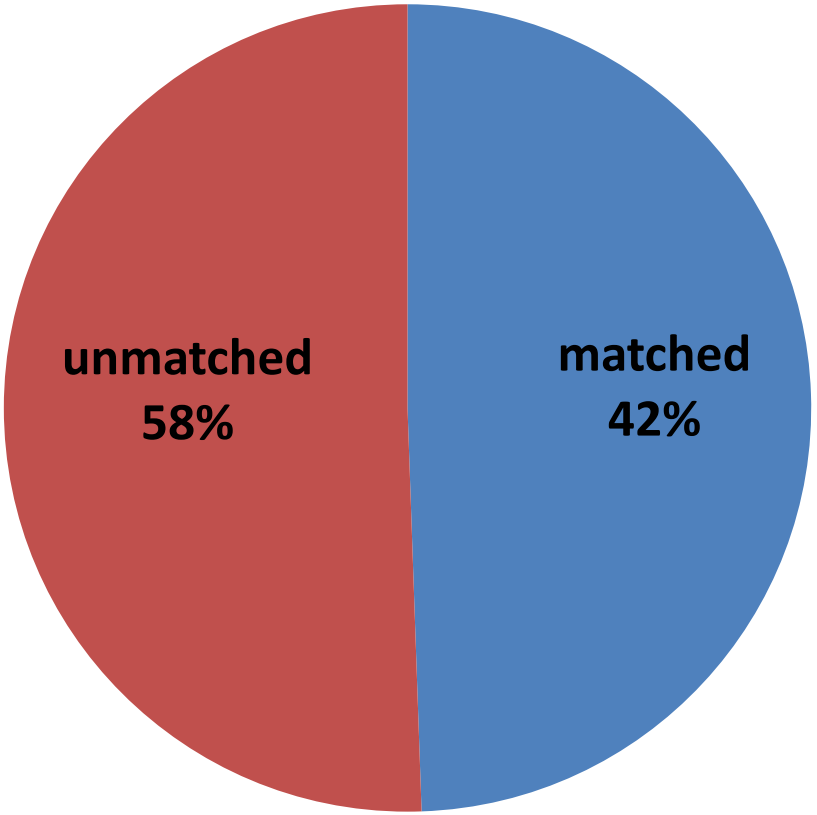
RP-RP分离



The data analysis strategy



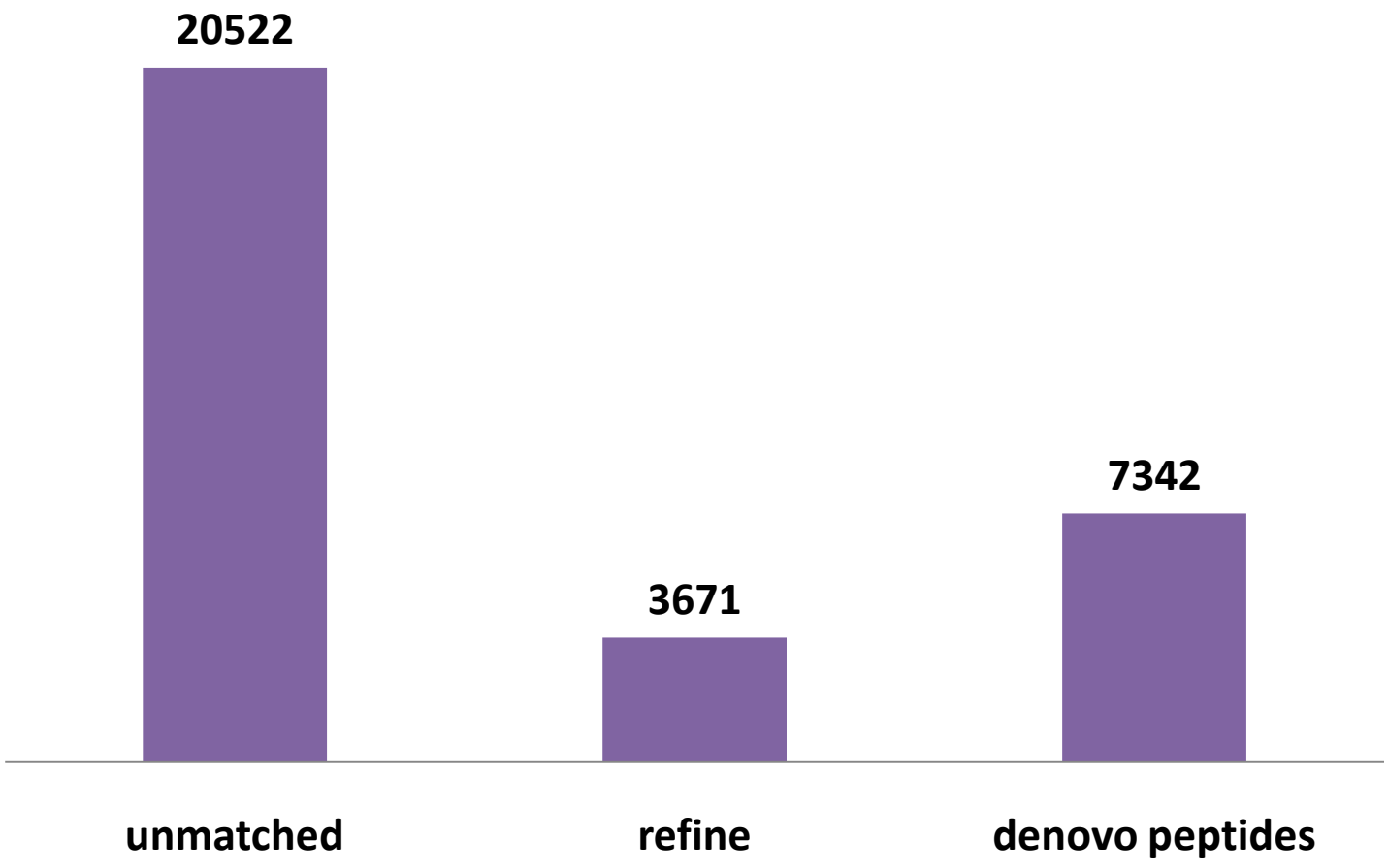
Unmatched spectra after Mascot search



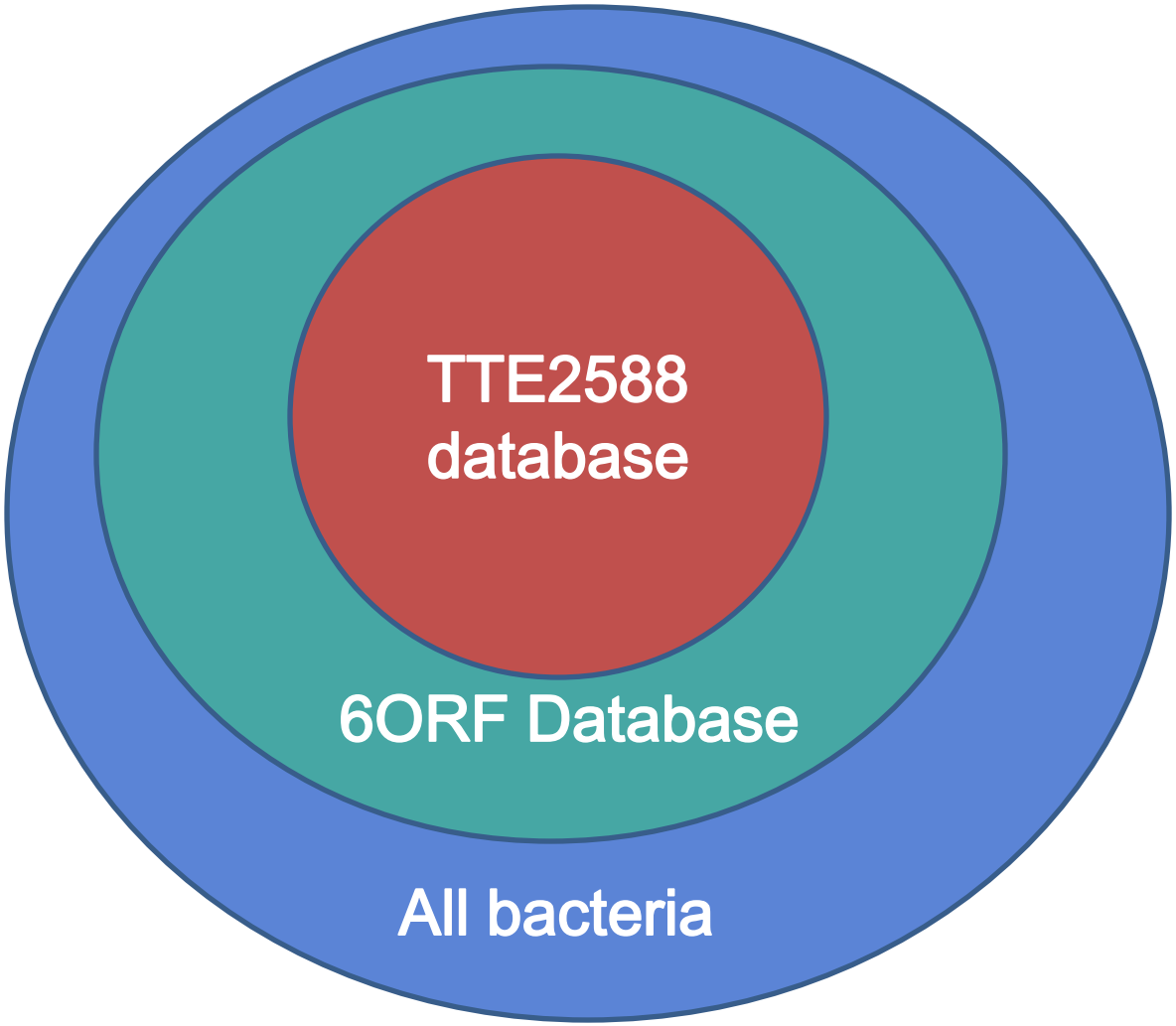
Spectra refine

- Merge the spectra with precursor mass difference no more than 2ppm and retention time difference no more than in 1min.
- The precursor charge state between 1-3.
- Precursor mass between 800-3000Da.
- Spectra quality value over 0.7.

Peptides obtained by *De novo*



Database homology search



Peptide and protein filter

Peptide

- Peaks evaluation score is over 0.8.
- Match to 6ORF database, at least 7 amino acid per peptide.
- No match to reverse database.

Protein

- At least 2 unique peptides per protein.
- At least 7 amino acid per peptide homology matched to the target .

Definition of new peptide and new protein

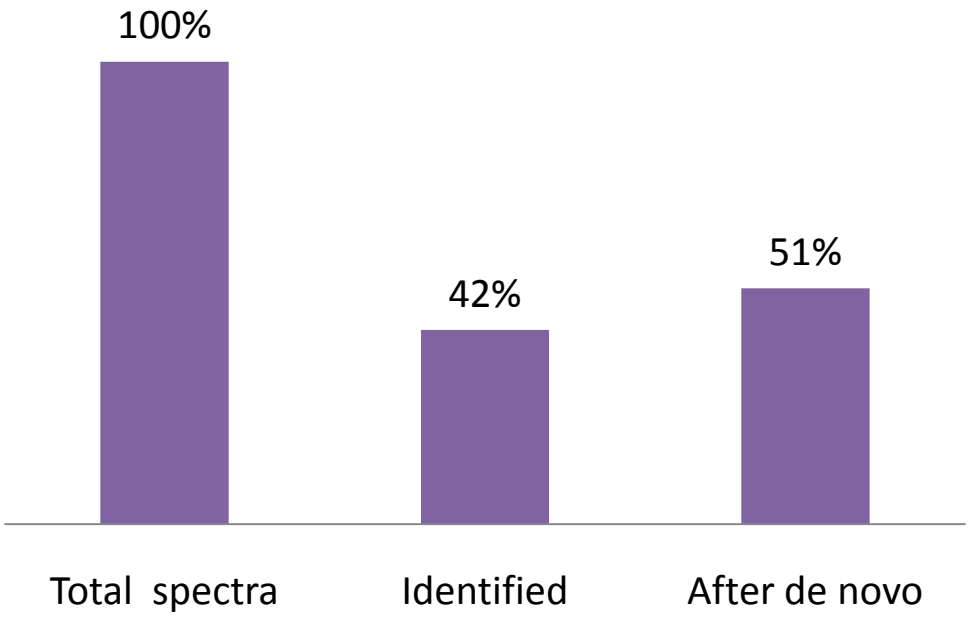
New peptide

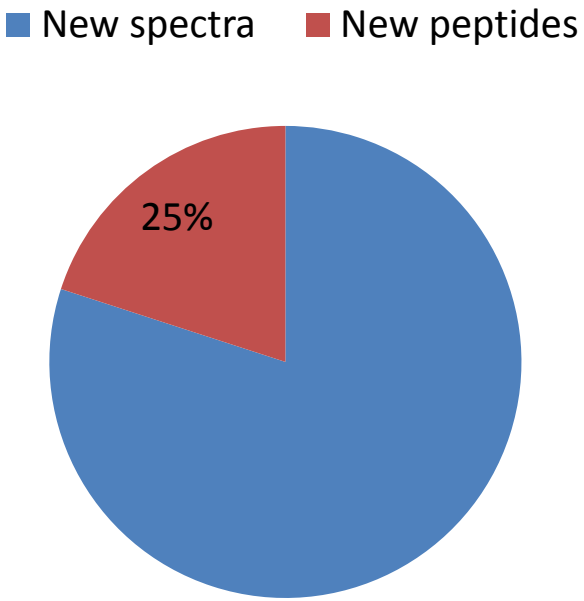
- Match to 6ORF database, but no match to TTE2588 database.
- No match to reverse database.

New protein

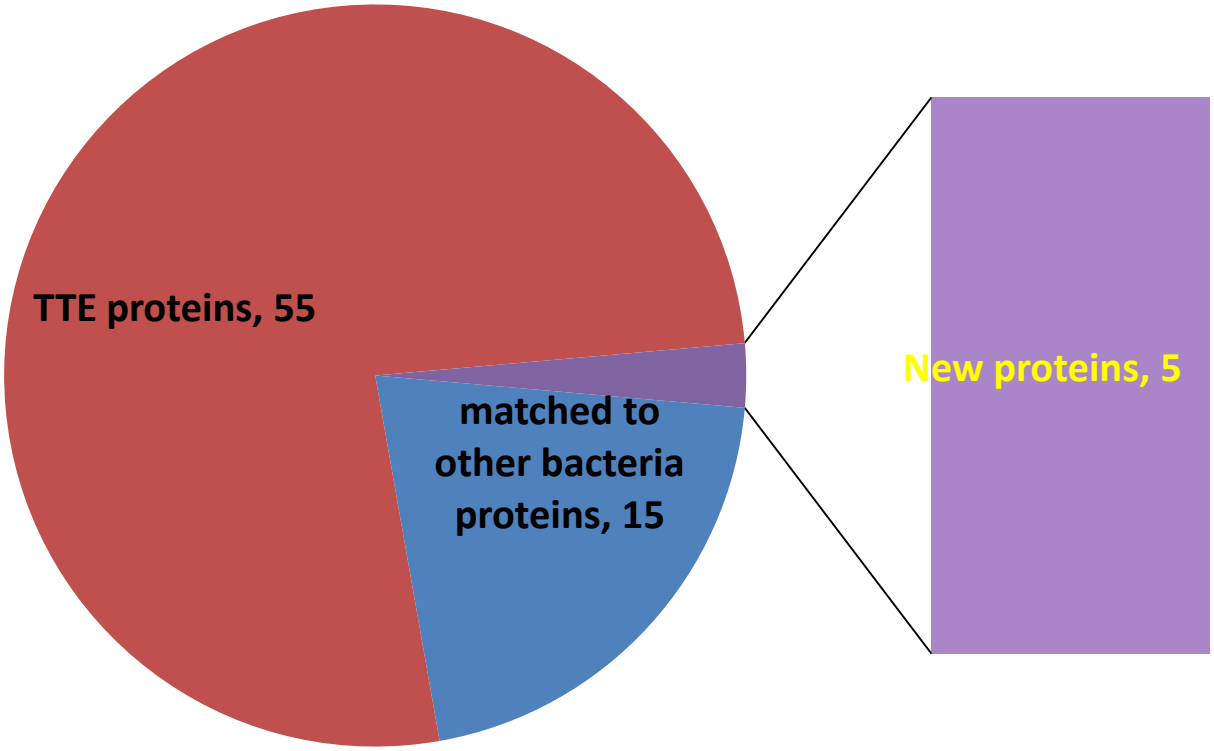
- Exist in 6ORF database, but not in 2588 database.
- Match to some other proteins by BLAST.

Peptide identification rate has improved combinning with *de novo*



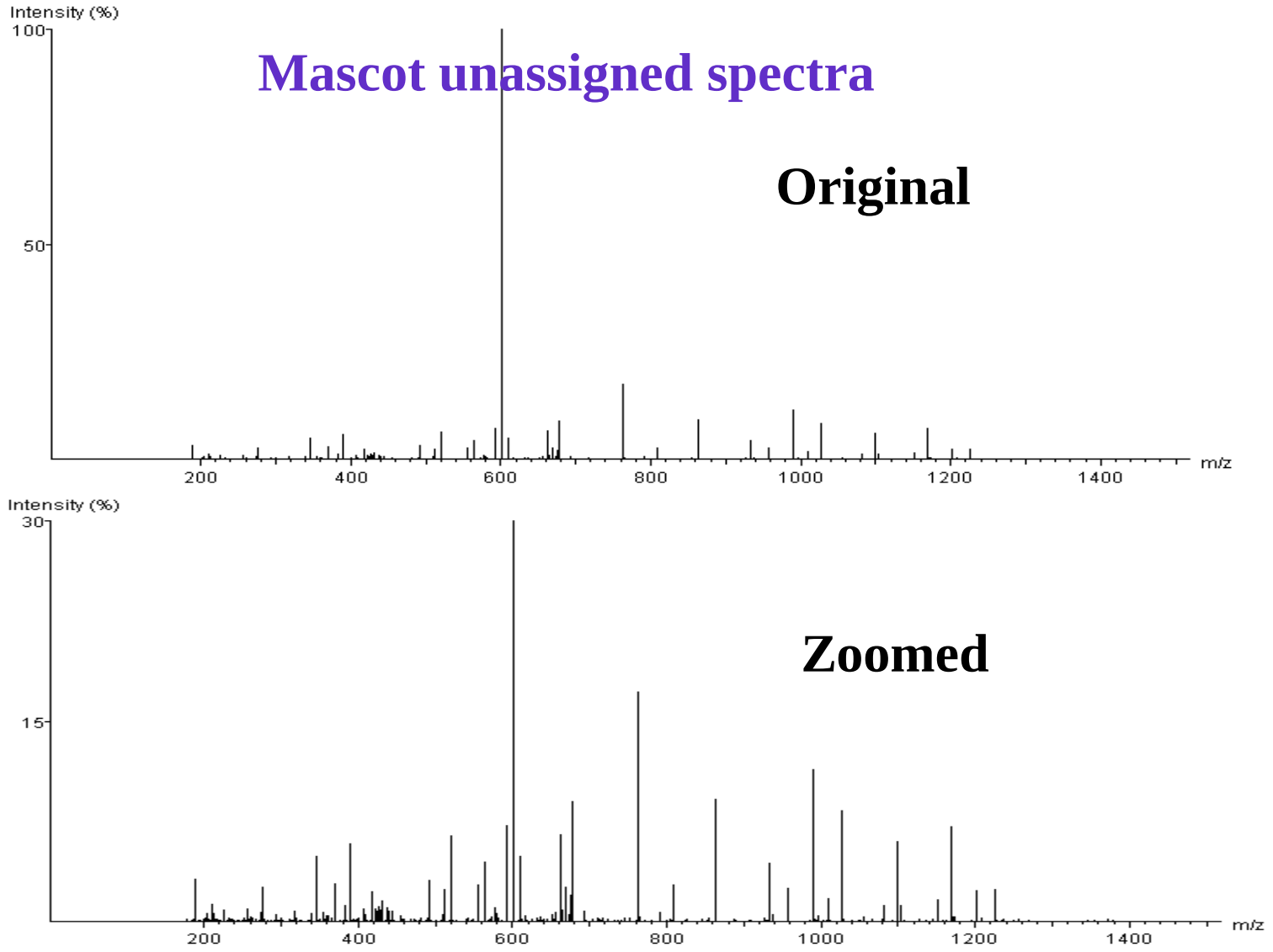


Proteins identified with *De novo* peptides

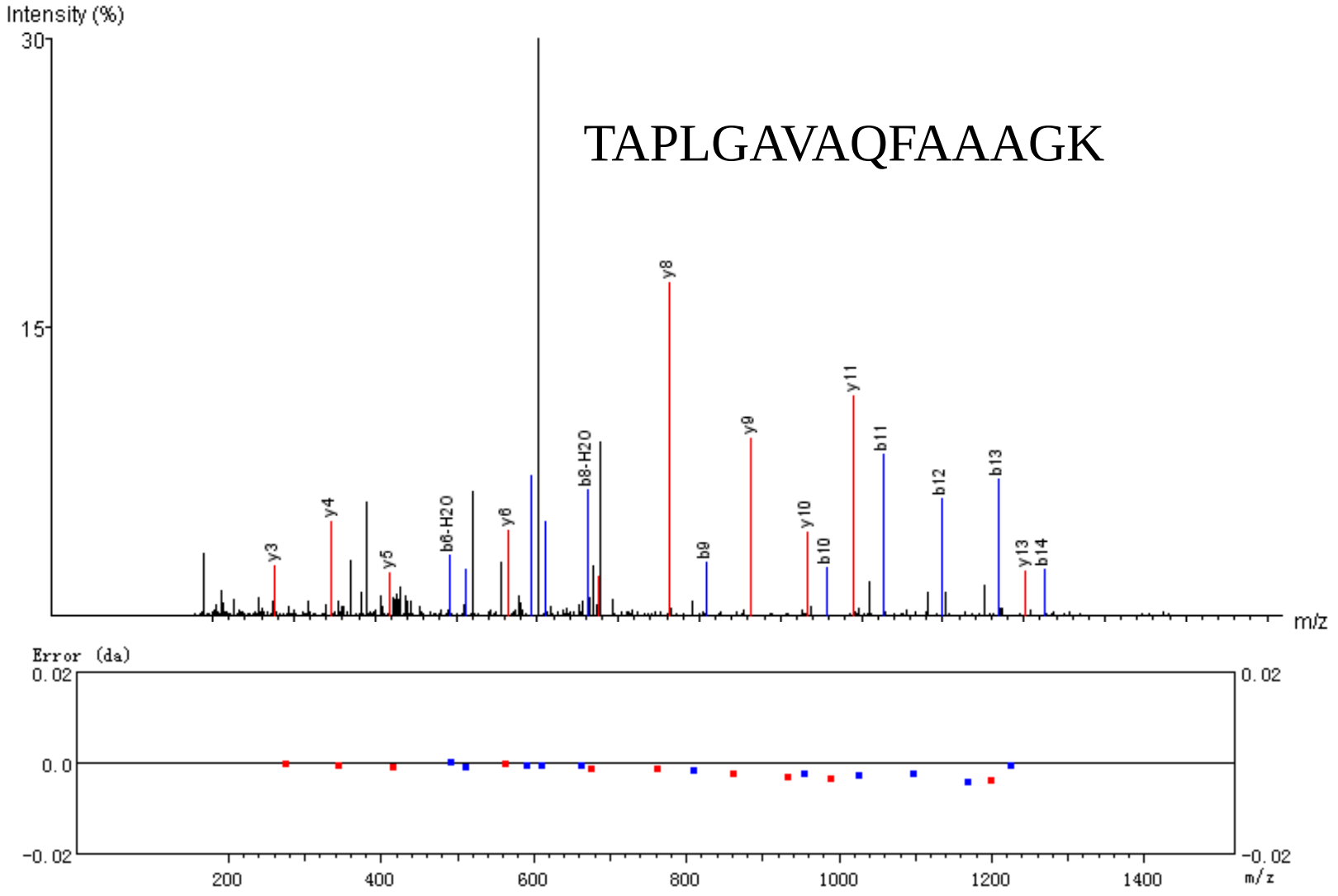


Totally 70 unique proteins were found with high confidence against the NCBI nr bacteria database

Peptide identification rate improved combinnig with *de novo*



After *de novo* sequencing



This spectrum matched to PFOR protein

De novo improves the reliability of identified proteins

```
  1  MK*IIVTEK*IS ENGIDYLKKY ADVDVKTNIS REELLEVIKD YDAIIVRSAT
 51  KVDRELIEKG EKLKVIGRAG NGVDNIDVEA ATQRGILVVN TPAGNTIAAA
101  ELTIGLMLAI ARNIPQAYHA ALNGDFRRDR FKGVELNGKT VGIIGLGRIG
151  SLVASRLAAF NMR**VIAYDPY MPDER**FEKCG VKRVTLDELL EQSDFITIHI
201  PKTEETKKMI GEKEFKKMKK GVRIVNAARG GIIDEKALYN AIKEGIVAAV
251  GLDVLEVEPK YNVEHQDFHN PLLELPNVVF TPHLGASTYE AQENISIAIA
301  QEVISALNGN LYGNIVNLPG LKSDEFSRLK PYMKLAEVLG ALYYQINETP
351  A*KLIEVIYR*G EVAKSNTEIV TLYAIKGFLK PILEEDVSVV NAKLRAKEMG
401  IEIVEGKIEE IDHYSSLVIL KITDTNGKRT QFAGTTYGEE LRIVEYMGHK
451  VNFEPTEYML FVKNKDVPGV IGHIGNVLGD FGINISTMQV SPNKNDGTAL
501  MLVSTDKEIP EEAVESLNKL NSIIKAKAVK GLV
```

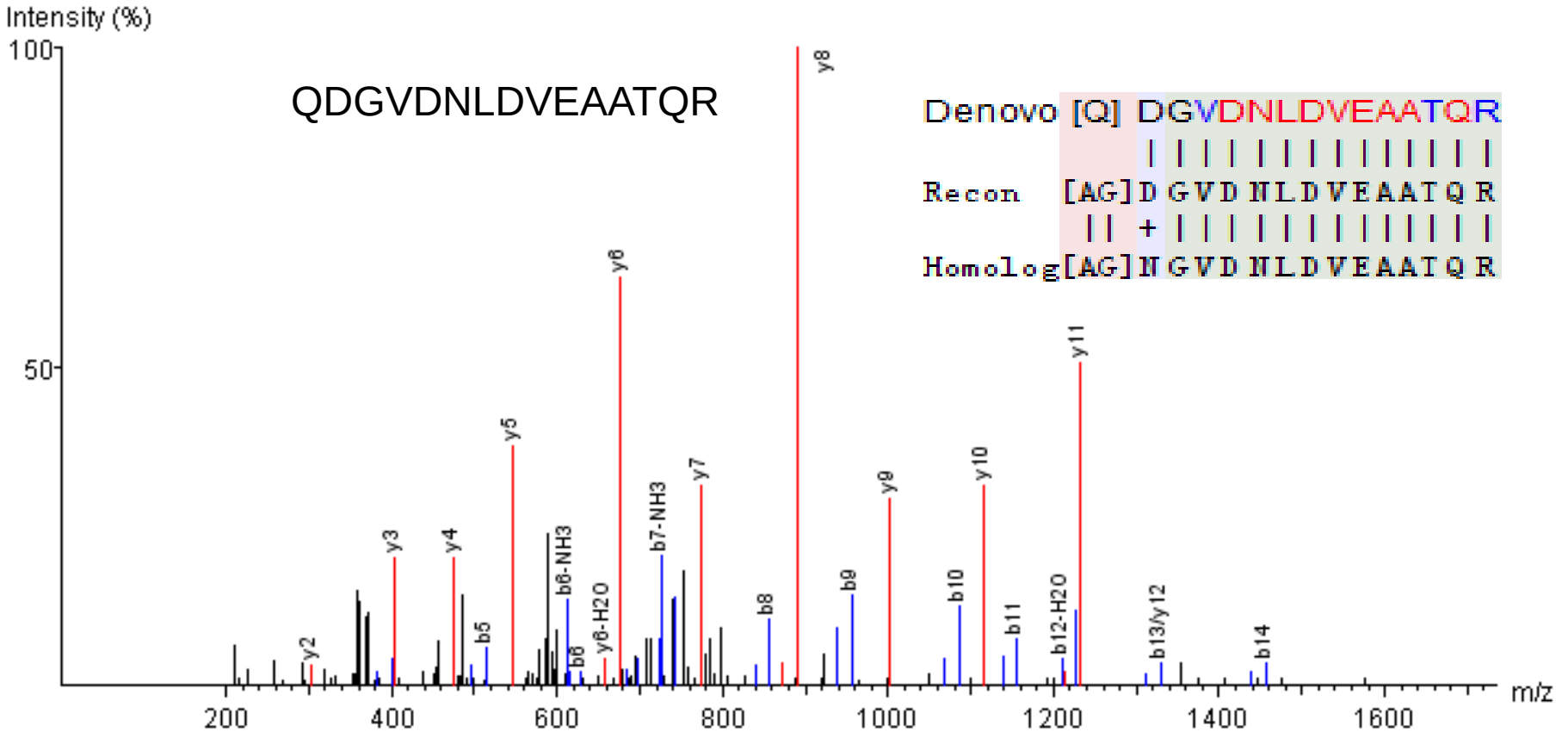
gi|20517622 Mass: 58877 Score: 25 Matches: 4(1) Sequences: 3(1) emPAI: 0.06

Phosphoglycerate dehydrogenase and related dehydrogenases [Thermoanaerobacter tengcongensis MB4]

Query	Observed	Mr(expt)	Mr(calc)	ppm	Miss	Score	Expect	Rank	Unique	Peptide
594	351.7234	701.4322	701.4323	-0.23	0	5	0.79	9	U	K.IIVTEK.I
3981	453.2757	904.5369	904.5382	-1.41	0	4	1.6	3	U	K.LIEVIYR.G 3980
20998	734.8439	1467.6733	1467.6704	1.99	0	25	0.0029	1	U	R.VIAYDPYMPDER.F

It is very difficult to determine if this protein was true or not

De novo improves the reliability of identified proteins

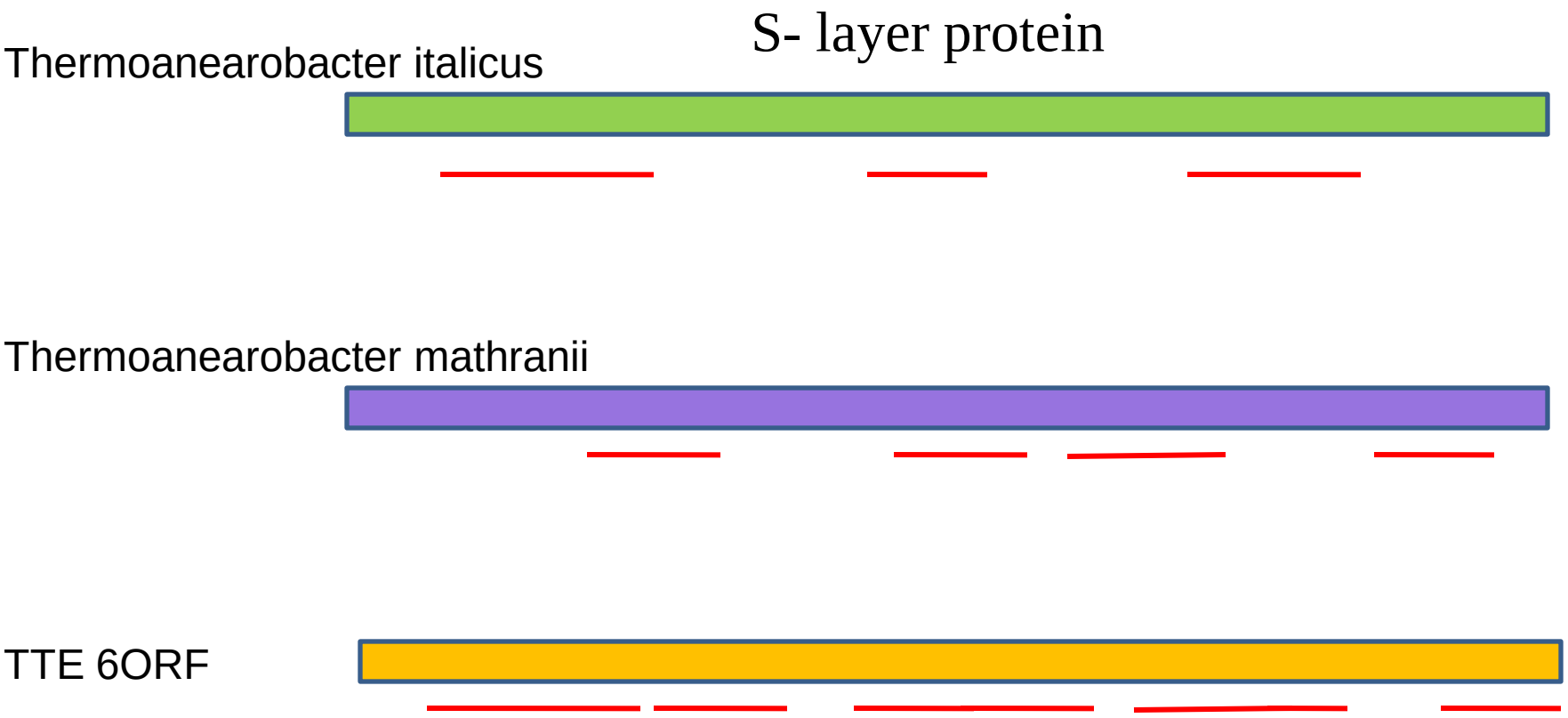


Mascot search combined *with de novo*

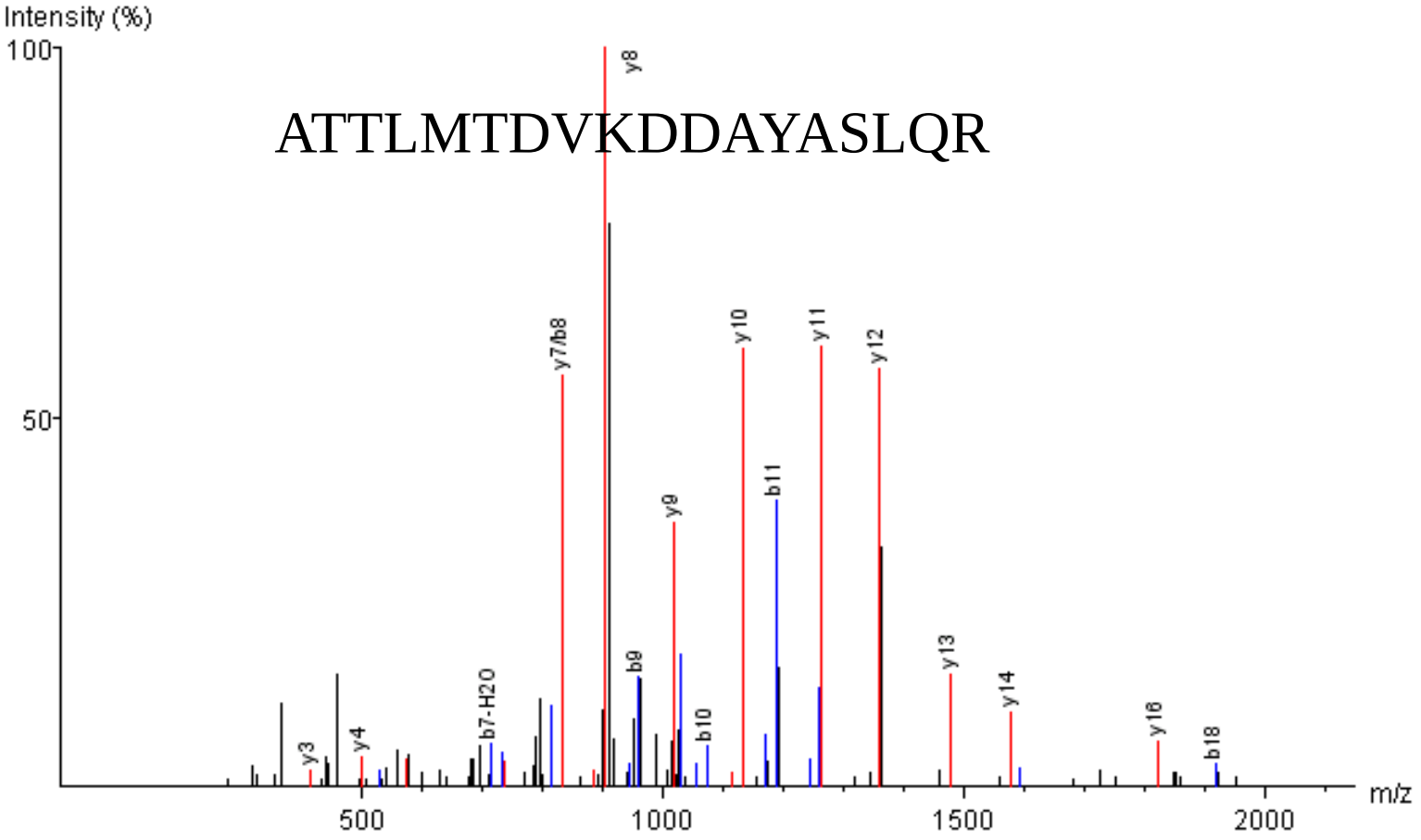
1 MK***IIVTEK***IS ENGIDYLKKY ADVDVKTNIS REELLEVIKD YDAIIVRSAT
51 KVDRELIEKG EKLKVIGRAG NGVDNIDVEA ATQRGILVVN TPAGNTIAAA
101 ELTIGLMLAI ARNIPQAYHA ALNGDFRRDR FKGVELNGKT VGIIGLGRIG
151 SLVASRLAAF NMR**VIAYDPY MPDER**FEKCG VKRVTLDELL EQSDFITIHI
201 PKTEETKKMI GEKEFKKMKK GVRIVNAARG GIIDEKALYN AIKEGIVAAV
251 GLDVLEVEPK YNVEHQDFHN PLLELPNVVF TPHLGASTYE AQENISIAIA
301 QEVISALNGN LYGNIVNLPG LKSDEFSRLK PYMKLAEVLG ALYYQINETP
351 A***KLIEVIYR***G EVAKSNTEIV TLYAIKGFLK PILEEDVSVV NAKLRAKEMG
401 IEIVEGKIEE IDHYSSLVIL KITDTNGKRT QFAGTTYGEE LRIVEYMGHK
451 VNFEPTEYML FVKNKDVPGV IGHIGNVLGD FGINISTMQV SPNKNDGTAL
501 MLVSTDKEIP EEAVESLNKL NSIIKAKAVK GLV

The reliability of this protein is obviously increased

A new protein



One of the peptides matched to s-layer protein




gi |289579402 MKSLKKLIAVVLTFALVFSAMAVGFAATTPFTDVKDDAPYASAVARLYALNITNGNTDGT
gi |255337715 MKNLKKLIAVVLTFALVFSAMAVGFAATTPFTDVKDDAPYASAVARLYALNITNGNTDGT
gi |289579402 YGVDQPVTRAMMvVFVNRLSGYRNLAEaAKNDaPAFsDVSKNYWAVGDINLAAKLGLTHG
gi |255337715 YGVDQPVTRAMMtVFVNRLSGYRNLAEmAKNDtPAFkDVSKNYWAVGDINLAAKLGLTHG
gi |289579402 VGNGkFNPEGKVTYAQALGFMLNALGYKDLSWPYGVLAKAQDLGLavvsDiglNDVInRG
gi |255337715 VGNGmFDPEGKVTYAQALGFMLNALGYKNLSWPYGVVAKAQDLGLtaglNrayNDVVtRG
gi |289579402 qLALIMDKALDQEVVkyYDeNGNPVLGDKLISKItDtTdYLIVATPDVDSSVAdGKVLVQ
gi |255337715 dLALIMDKALDQQIVtsYDtNGNPVLGNKLISKVaDvTrYLVVATPDVDSNVAqGKVLVQ
gi |289579402 eVastSTtGVrSFKtATTIDAGdIDFNQYLGKVVtIYTaKnGDePLAVDVVTTDyTFTAn
gi |255337715 gIkdvNSdGViTFKaATTINAGtVDFNKYLGKVVdVYSiKgGD-PVSVDVVSTDkTFTAk
gi |289579402 NdNnVANAVYDEdgnyIEL--sSktpIVYNGvKTTLgadgvvIYDGANVtLTDTDNDGtY
gi |255337715 SfNvVSNSVYNDgskvVDIdtpAnvtVIYNGgKTTLdqvatkVYDGANVaLTDTNNDGkY
gi |289579402 DYAVVTnAFKygpLtVeSDVsASDaYIktNgV---SlQVSGgsIdkVVVTGSVSkLSDIE
gi |255337715 DYAVITgAYK-asVvVtADVkSTDkFLnvNnVynsSyRIAGdpVktVVVTGSVTsLTDIK
gi |289579402 tGDVVYYAvSaDGSKVTLfVIRDsItGEVTKVAqasdgTyTvTIdgEDYeVS----GNyT
gi |255337715 aGDVVYYAsTiDGSKVTLlVVRNkVeGKITKVA-ydgsTtTaTIgdKDYtVAqyinGNaS
gi |289579402 ---pqVGdEGTFaLDKDGkIaGFIGvtATeNYAIVLgiD-DDssaNpQIKLfTSEGKtvI
gi |255337715 gakatVGqEGTFvLDKDGnIiGFIGkqATaNYAVLLafNaNDvwnNgKVKLlTADGKvnV
gi |289579402 YpydtSgdipavgDLISYSLDSDNtVTdItvygnknDSpN------dwgYDsDTYVLa--
gi |255337715 YsttvTsttygtnDIITYSIDSNNvVTaIns-pktaDTdNvvainasaaYNkDTHVLtvs
gi |289579402 -daYYLdSSTVVFNvy--DDdYTvVdVSDITvDSLNVvaMakDdYGnVEALvIDEASLS-
gi |255337715 gttYYMtNNTVIFNynstDDsYStVkLSDITkDTLNVkqLvaDqYGyVKAIyLNESSLTa
gi |289579402 EseeASvLYGiVTdySTVktsdG-TyYKItVLaNnAEQTFTTttDVakPvkSTetSvtvy
gi |255337715 EqsvSStVYGyVTgvTTIdlgsGnTqYKLnVLvNgSEQTYTTkvNLt-PtpSTgaAa---

Why could not find the homology protein in *TTE*?

LIAFVVSFALVFGAMAVGFAATTPFTDVKDDAPYASAVARLVALDITKGVGDGKFGVDQP

VTRAQMVTFVNRMLGYEGLAEMAKAEKSVFKDVPQNHWAVGHINLAYQMGIAKGVGD

GKFDPNGRLTYAQALAFVLRALGYQDLSWPYGVLAKAQDIGLTAGINLAYNQVMLRGDL

ALVLDRALQTPMVKYVDGKETQGDKLISKVANVTKYLVVATPDVDSNVAAGKVQVKGIK

EVKDGVITFADATTINAGNIDFNKYLGKVVEVYTVKSTGEPVFVDVTATPEKSFTAKRFEV

VNTVVYNDNKKVVEIPSPTEVTVIYNGGKTTLDQVLAKAKVEDGASVTILAPDNKTYNY

MIVNDSFKYQNVRVTADVKAGDKFINANSSLRIAGDPVKTVIVKGSVDKVTDIKANDIIY

YGTTVDGSKVTILVVRDKVEGKVTTVIDDGAKVVINDKTYTVKGYAENKTPAVGDEGVF

VLDKDGNIVYAILKVSAAAENYAIALAAEAYGPLTGGKVELLTAEGKKILDAKYDAAVAA

ETYARNKDLVTYTVKDNKVDSVKRVE

By 6 ORF database search

Why could not find the homology protein in *TTE*?

Shift frame

Summaries

Spectra identification rate improved combined the database search with *de novo* sequencing.

The matched proteins are mostly *TTE* proteins, which indicate the homology search results are highly credible.

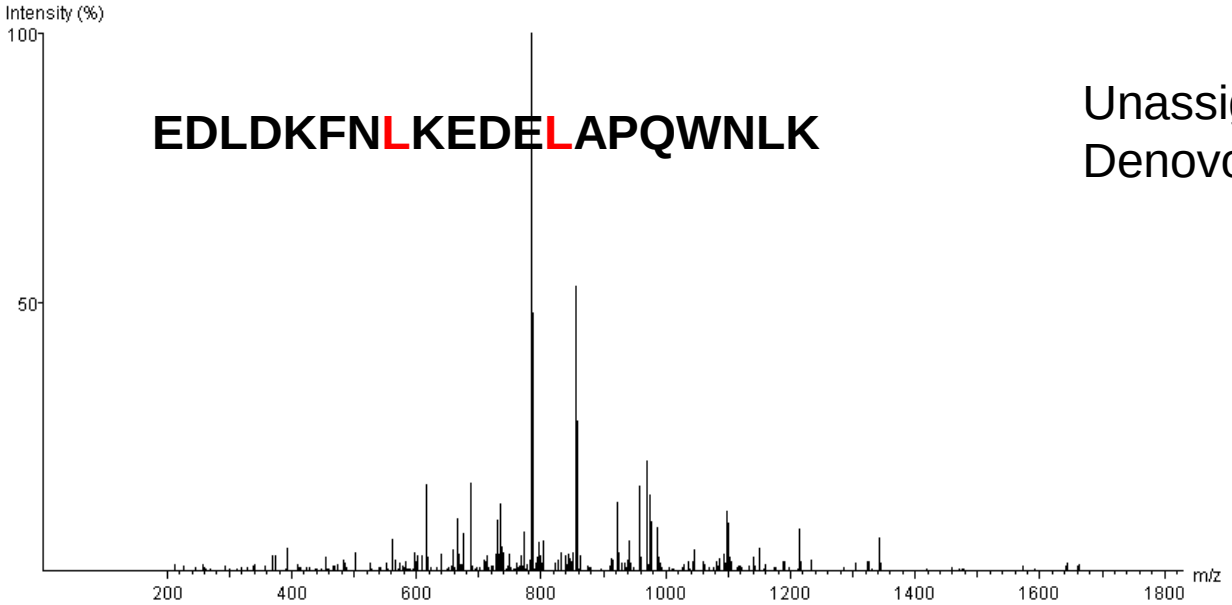
Most of the proteins matched to other bacteria are also exist in *TTE* as indicated by BLAST.

There are 2 proteins matched to other bacteria but not found being exist in TTE, which indicating TTE genome miss-annotation.

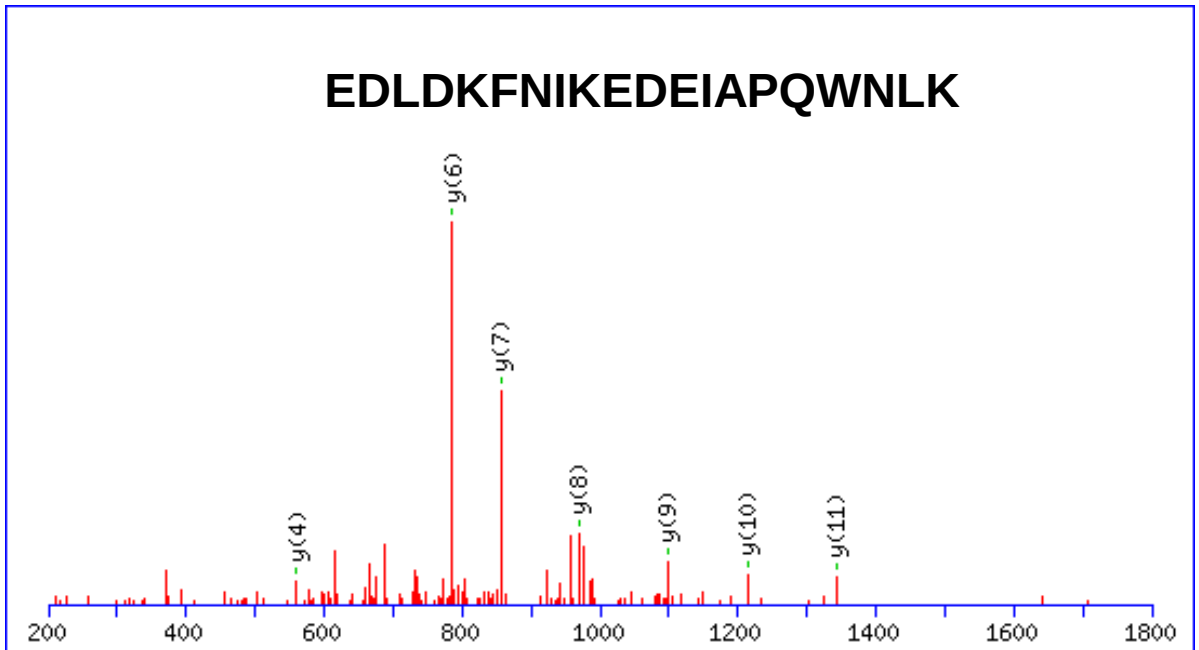
conclusions

- proteome could be performed by *de novo* even under unavailable of the genomic data.
- Spectra identification has improved although the protein identification increased unobviously.
- *De novo* sequencing could help correct the genome annotation.

Discussion



Unassigned
Denovo by PEAKS



Mascot
matched

Acknowledgement

We appreciate the effort contributed from the research group of bacterial proteomics in Beijing Genomics Institute, CAS.

We thank the 973 grant support from Department of Scientific Technology, China.

We thank for CNCP meeting.