International Journal of Advanced Computer Technology (IJACT)
ISSN:2319-7900

Evaluation and Comparison of Aligners for De Novo Sequencing

Simin Zhu; Huamei Li

Abstract

In high-throughput proteomics research of tandem mass spectrometry, de novo sequencing provides a novel method to interpret MS/MS data without any help of sequence database and discover new organisms. In this paper, we have systematically evaluated and compared the capability of mainstream de novo sequencing software via testing data sets which have been correctly identified by Mascot and Sequest, so we can intuitively find out the optimal de novo sequencing software for protein identification.

Introduction

In recent years, tandem mass spectrometry has got a great development in the field of proteomics. In a typical LC-MS/MS experiment[1, 2, 4], thousands of MS/MS spectra can be generated, how to interpret LC-MS/MS data rapidly and efficiently is still a challenge, although many peptide identification algorithms have been proposed[3, 5, 9, 10].

Database search and de novo sequencing are parallel methods for protein identification, the former is the most popular approach, which basic thought is to score the mass spectra against a database of all candidate peptides to detect significant matches[4, 6, 9, 15], such as Mascot, Sequest, SQID, OMMSA, ProVerB and pFind. However, the basic idea of the latter is to reconstruct the original peptide sequence without knowledge of genomic sequences or organism, including Lutefisk, PepNovo, pNovo, SeqMS and Peaks[3, 5, 6, 12]. Database search assumes that the genome are accurately sequenced and annotated which are hardly satisfied. De Novo sequencing provides a novel way to identify the MS/MS data only according to spectra itself[14].

De novo sequencing contains four mainly procedures: construct the graph of MS/MS spectra, confirm the ions type, establish sequencing algorithm and scoring model[3-6]. How to contruct spectrum graph and determine ions type are the fundamental of de novo algorithms, sequencing algorithm provides searching methods for spectrum graph, scoring function is the heart of de novo algorithms, varies protein identification algorithms have different score function, there are often determine the advantages or disadvantages of the algorithms[4, 8, 14].

Here, we have systematically recommend the mainstream methods of sequencing algorithms and the categories of scoring model. Meantime, in order to evaluate and compare the sensitivity of widely used de novo sequencing software, we utilize the testing dataset which correctly identified by commercial database search software Mascot and Sequest[7, 11, 15]. We hope our research can provide a reference for researchers.

1 De Novo Sequencing Algorithms

1.1 Constructing The Spectrum Graph

An alternative approach of de novo sequencing is to transform the spectra data as a directed a cyclic graph, where a node corresponds to a mass peak and an edge is connected by two nodes which differ by the total mass of an amino acid. For a spectrum $S=\{s_1, s_2, \cdots, s_m\}$, the corresponding mass is $M=\{w_1, w_2, \cdots, w_m\}$, the parent mass of S is W, the spectrum graph can be constructed follows:

- Generated two vertexes, denoted as v_0 and v_m, which mass are 0 and $W-18$, respectively.
- The type of fragment s_i is N-terminal or C-terminal can not be determined, so it is necessary to create complementary $\bar{s}_j$, which mass are w_j-1 and $W-w_j+2$, respectively.
- Assume the differ mass by two fragments denoted as : $|w_i-w_j|\le\Delta m$, meantime, the mass of amino acid $AA=\{aa_1, aa_2, \cdots, aa_{20}\}$, δ denoted as the allowable error of instrument. If $\Delta m\le\delta$, connecting the two fragments (i.e. two nodes) by an edge.

1.2 Confirming The Ions Type

How to confirm the ions type is crucial for de novo sequencing, in a general way, there is not enough information to confirm the ions type. Hence, we need to find out immonium ions in the region of the lower mass, then confirm the other ions type. SHERENGA utilize the offset frequency function to confirm the ions type. PRIME use graph theory to distinguish the ions type.

1.3 Sequencing Algorithm and Scoring

The de novo sequencing problem is to reconstruct the

ISSN:2319-7900

peptide sequence from a given tandem mass spectral data, given a spectrum graph, our aim is to find a maximum score path for s_0 to $\bar{s}_0$. A number of de novo sequencing algorithms have been reported for deduction of protein sequences from MS/MS spectra data. Although several instrument manufacturers have developed, they are not satisfy for proteomics research. Most of software packages, including Lutefisk, PRIME, SHERENGA, and SeqMS use dynamic program to search spectrum graph.

Scoring function is the core of the de novo sequencing algorithms, varies algorithms have different scoring model. In order remedy the disadvantage of missing peaks, Peaks provides a new mathematics model to compute the reward/penalty that a y (or b) ion has mass m , the formula as follows:

$$f\left(\frac{h_1}{h}\right)\times f\left(\frac{h_2}{h}\right)\times f\left(\frac{h_3}{h}\right)\times \exp\left(-\left(\frac{m'-m}{\delta}\right)^2\right)\times \log h$$

Where:

m' = the mass of a $y-ion$

m' = the mass of the observed peak for that $y-ion$

δ = the mass error tolerance of the spectrometer

h_1, h_2, h_3, h_4 = the relative abundances of the observed $y-ion$ peak and the corresponding x , $y-H_2O$, $y-NH_3$ peaks

The approach of Peaks is different from the spectrum graph model used by previous algorithms, which consideration is attempt to give a reward/penalty score for improving the accuracy of the de novo sequencing results. PepNovo proposed a probabilistic network model which attempted to interpret the MS/MS data, the algorithms can score any number of peptides, the heart consideration is to utilize the likelihood ration hypothesis test to determine whether the peaks observed in the mass spectrum are more likely to have been produced under its fragmentation model than under a model that treats peaks as random events. The mainly scoring function as follows:

$$Score(m,S)=\log\frac{P_{CID}(\vec{I}\mid m,S)}{P_{RAND}(\vec{I}\mid m,S)}$$

The details of the parameters reported by ref 4, if the $Score(m,s)>0$, the denotes generated by actual peaks, otherwise, generated by random peaks[4].

on the whole, all of de novo sequencing algorithms are according to the quality of parent ions, enumerate parts of possible candidate peptides, then compare the candidate peptides by scoring function, finally, find out the best matching of candidate peptides.

2 Searching and Comparison

In order to validate the accuracy and effectiveness of mainstream de novo sequencing software,including Lutefisk, Peaks, PepNovo, pNovo+. We utilize 18 MS/MS spectra from Micromass/Waters QTOF Ultima and LTQ-Orbitrap instruments, which correctly identified by Mascot and Sequest. The parameters set as follows:

- Micromass/Waters QTOF Ultima
  - Parent ions tolerance error : 0.2 Da
  - Fragment ions tolerance error: 0.5 Da
- LTQ-Orbitrap
  - Parent ions tolerance error : 10 ppm
  - Fragment ions tolerance error: 0.5 Da

The specific identifying results as the following table.

Table1 The results of de novo sequencing search

Spectra	Sequence	Peaks	pNovo	PepNovo	Lutefisk
QT20060328_Den18mix_06.2112.2112.2.dta	DGPLTGTYR	MHNTYSVK	YMSTPPQR	None	[382.14]T[WY]K
QT20060328_Den18mix_03.2565.2565.2.dta	AWMSAAAIAK	FDPLADWR	MMDFHNPK	None	[262.10]PLAD[RW]
QT20060328_Den18mix_02.2326.2326.2.dta	VKVNSVLM*K	QVVGVDC(+57.02)TR	QVVGVDAFAK	None	VQVGVD[261.11]R
QT20060328_Den18mix_02.3025.3025.2.dta	□ASGGAAVSISIK	GTGWNLPMGK	TGWGNNGLGGQ	None	[158.07]GWNLPCR
QT20060328_Den18mix_04.3013.3013.2.dta	SAGTNTVASLR	ASGSYYFGPK	TGGAPCNTGNGK	None	DQGT[227.17]FPGK
QT20060328_Den18mix_01.2517.2517.2.dta	VAAAFPGDVDR	VAAAFLLAADR	VAAAETDVWR	None	[170.11]AAF[198.10]AGANK

ISSN:2319-7900

QT20060328_Den18mix_10.2121.2121.2.dta	SRGRGGGGGGFR	SSKLTSLAYK	GVEHYEGSDK	None	[SS][335.11]PNVN[AP]
QT20060328_Den18mix_06.2070.2070.2.dta	LALGASTGLGLR	LAANESKGPNK	LATQTNMPPR	None	LAANEY[EY]R
SXS_D39_090104_0.4351.4351.1.dta	NVALAVK	NVALAVK	KVGLLGAG	VYPVALAAR	NVALAVK
SXS_D39_090104_0.2235.2235.2.dta	VNAEAGLK	VNARWR	SSGLVPNK	VNDANLR	VNATKLR
SXS_D39_090104_0.4326.4326.2.dta	NLMPNPK	NLMPPNK	ARARDPGA	LNNLVLK	VQ[227.12]VLK
SXS_D39_090104_20.4594.4594.2.dta	VGINGFGR	VGLNGFGR	PASGKGFR	VGLNGFGR	VGLNGFGR
SXS_D39_090104_20.1994.1994.2.dta	YGGLGTQK	YGGLGTQK	GTFGKGEK	YGGLGDNK	YG[212.15]LMK
SXS_D39_090104_0.1574.1574.2.dta	QATPGGAVK	FFHLHK	KPPNSSAK	YETNDGK	[170.11]R[NS]GPK
SXS_D39_090104_20.6290.6290.2.dta	EITGLGLLK	ELTGLGLK	GKMAKAPK	ELTGLGLK	NKT[170.11]GLK
SXS_D39_090104_0.3891.3891.2.dta	LDVEASAK	LDVEASAK	LGASDLEK	LDVEASAK	LDVEASAK
SXS_D39_090104_0.3000.3000.2.dta	YDTTQGR	YDTGASAR	YDAMKGGV	DYTGWAK	[278.09]TG[229.11]R
SXS_D39_090104_0.5561.5561.2.dta	LSGGVAVIK	LSGTKPLK	SLGGKLLVG	GNLAVEK	LSR[NP]TR

3 Discussion

In this paper, we can see that PEAKS and PepNovo are performed more accuracy and robust than other software below LTQ-Orbitrap instrument. However, we also can find almost all spectra generated by Micromass/Waters QTOF Ultima instrument has lower quality and the searching results showing multifarious. Although de novo sequencing algorithms still beset with difficulties, the technology of instruments have been develop rapidly, and the research of protein identification algorithms become more thorough, we have hope of de novo sequencing algorithms.

References

[1] Diaz E, Machutta CA, Chen S, Jiang Y, Nixon C, et al. Development and validation of reagents and assays for EZH2 peptide and nucleosome highthroughput screens. J Biomol Screen 2012, 17: 1279–1292.

[2] Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. Method to Correlate Tandem Mass-Spectr of Modified Peptides to Amino Acid Sequences in the Protein Database. Anal. Chem. 1995, 67 (8), 1426–1436.

[3] Zhang, L.; Reilly, J. P. Peptide De novo Sequencing using 157 nm Photodissociation in a Tandem Time-of-Flight Mass Spectrometer. Anal. Chem. 2009, 82, 898–908.

[4] Siuzdak,G.The Expanding Role of Mass Spectrometry in Biotechnology; MCC Press: San Diego, CA, 2003.

[5] Taylor, J.A, and Johnson, R.S. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrom etry. Rapid Commun. Mass Spectrom. 11, 1067–1075.

[6] Yates, J.R., GrifŽ n, P.R., Hood, L.E., Zhou, J.X. 1991. Computer aided interpretation of low energy ms/ms mass spectra of peptides, 477–485. In Villafranca, J.J. ed., Techniques in Protein Chemistry II. Academic Press, San Diego.

[7] Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A. 1999.˘De novo peptide sequencing via tandem mass spectrometry. J. Comp. Biol. 6, 327–342.

[8] Han, Y.; Ma, B.; Zhang, K. SPIDER: Software for protein identification from sequence tags with de novo sequencing error. J. Bioinform. Comput. Biol. 2005, 3, 697-716.

[9] Taylor, J. A.; Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. 1997, 11, 1067- 1075.

[10] Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. NovoHMM: A Hidden Markov Model for de novo peptide sequencing. Anal. Chem. 2005, 77, 7265-7273.

[11] Allmer, J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. Expert Rev. Proteomics 2011, 8 (5), 645 − 657.

[12] Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20 (18), 3551 −3567.

[13] Wang, L. H.; Li, D. Q.; Fu, Y.; Wang, H. P.; Zhang, J. F.; Yuan, Z. F.;Sun, R. X.; Zeng, R.; He, S. M.; Gao, W. pFind 2.0: a software package for peptide and

protein identification via tandem mass spectrometry. Rapid Commun. Mass Spectrom. 2007, 21 (18), 2985–91.

[14] Chen T, Kao MY, Tepel M, etal, A dynamic programming approach to de novo peptide sequencing via tandem mass spectrumetry.[J] J Comput Biol,2001,8(3):325-337.

[15] Ma B, Zhang K, Hendrie C, etal, A powerful software tool for the de novo sequencing of peptides from MS/MS data [J].Rapid Commun Mass Spectrum,2003,17(20): 2337-2342.

Biographies

FIRST A. S.M Z Studying for a master's degree at Yunnan Minzu University, Kunming, Yunnan province. Currently, her research areas include protein identification algorithms. Contact email address: zhusimin2013@163.com .

SECOND A. H.M L Studying for a master's degree at Yunnan Minzu University, Kunming, Yunnan province. Currently, his research areas include protein identification algorithms and protein microarray data analysis. H.M L is the corresponding author, and his contact email address: li_hua_mei@163.com.