A Top-Down Driven Approach to *De Novo* Sequencing of Proteins

Kira Vyatkina[1], Lennard Dekker[2], Si Wu[3], Martijn Vanduijn[2], Vitalii Demyanyuk[4], Xiaowen Liu[5], Mikhail Dvorkin[1], Sonya Alexandrova[1], Theo Luider[2], Nikola Tolić[3], Ljiljana Paša-Tolić[3], and Pavel A. Pevzner[1,6]

[1] Saint Petersburg Academic University, RAS, Russia; [2] Erasmus MC, Rotterdam, The Netherlands; [3] Pacific Northwest National Laboratory, WA, USA;
[4] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics, Russia; [5] Indiana University, Indianapolis, IN, USA; [6] University of California, San Diego, CA, USA

Overview

In this work, we introduce the concept of a *T*-Bruijn graph and propose a fast and reliable approach, named *Twister*, to *de novo* sequencing of proteins from combined sets of top-down and bottom-up mass spectra using such graphs.

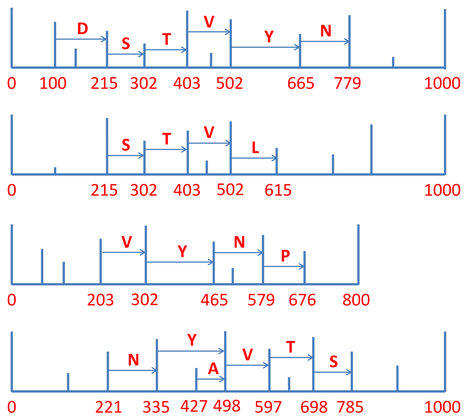
Introduction

Mass spectrometry is the most powerful technology for protein identification and characterization. The bottom-up strategy, which analyses peptides resulting from enzymatic digestion, and the recently emerged top-down strategy, which analyses intact proteins, are often viewed as complementary; however, when they are applied together, top-down techniques usually play a secondary role of a validation tool.

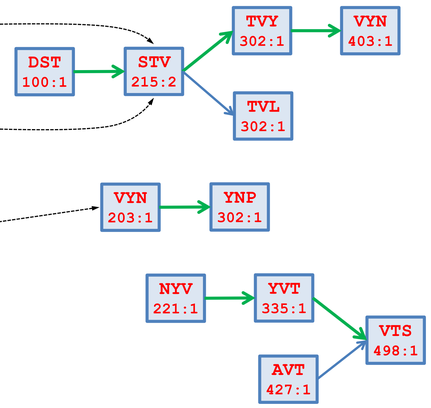
We describe a novel approach to *de novo* sequencing of proteins from combined sets of top-down and bottom-up tandem mass spectra. Thereby we grant the top-down data a leading role, and retrieve from it a number of seed fragments of the protein sequence, which then iteratively get extended, corrected, and whenever possible, merged with the aid of the information derived from the bottom-up dataset. In addition, the bottom-up spectra are handled in much the same way as the top-down ones. Finally, we point out that in case of contaminated samples or protein mixtures, our method outputs sequence fragments for various compounds, thus providing extra opportunities for analyzing a sample even in the absence of *a priori* knowledge on its composition. Our algorithms are implemented in a software tool *Twister*.

Methods

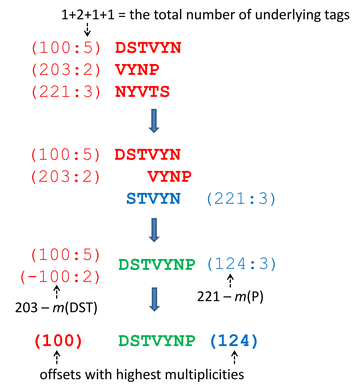
1) Tag generation



2) *T*-Bruijn graph construction and optimal paths extraction



3) Aggregated paths formation



$PM_{DSTVYNP}$ = **100** + *m*(**DSTVYNP**) + **124** = 1000: precursor mass estimate

When processing bottom-up mass spectra, we only compute optimal paths in a *T*-Bruijn graph and derive the amino acid strings they spell out. Those strings are subsequently used to validate, correct and extend the aggregated paths generated from the top-down data, to which end the former are appropriately aligned against the latter.

Results

Benchmarking: we tested *Twister* on a combined dataset for the alemtuzumab light chain.

Top-down experiments: a mixture of two antibodies, alemtuzumab (MabCampath) and adalimumab (Humira), was analyzed using an RPLC system coupled to an LTQ Orbitrap Velos with ETD and a Q-Exactive with CID or HCD, and 1523 ETD, 1282 CID and 1720 HCD spectra were acquired. Next, we applied the tool MS-Align+ to find the best protein-spectrum-match for each spectrum and the heavy and light chain of either antibody, and separated the 1,840 spectra assigned to alemtuzumab light chain into the top-down dataset.

Bottom-up experiments: alemtuzumab was digested with trypsin, chymotrypsin, proteinase K or pepsin, and analyzed using a nanoLC system coupled to an LTQ Orbitrap Velos with HCD; in total, 98,787 mass spectra were generated from both chains, and collected without separation into the bottom-up dataset.

Aggregated paths: obtained **41** in total, **16** of which passed the validation procedure.

Precursor mass estimate: the largest group of close values of PM estimates induced by the 41 aggregated paths comprised 7 values ranging from **23,553.837Da** to **23,554.911Da**. To obtain the final PM estimate for alemtuzumab light chain, we needed to subtract from their average of **23,554.455Da** the mass of **-2.016Da** corresponding to a disulphide bond formation, since most often, complementary ions correspond to a pair of polypeptides together containing precisely one disulphide bond. The resulting value of **23,556.471Da** differs from the theoretical mass of **23,566.703Da** of the alemtuzumab light chain by **0.232Da**, which is within **10ppm** of the latter.

Contaminants: the 9th top scoring aggregated path (which did not pass the validation procedure) had the amino acid sequence DFLQKKVAVLEDALEQ. A BLAST search for its reversed copy QELADELVAVKKQLFD against the non-redundant database reported a single sequence without mismatches (up to substituting L-7 with I) – and namely, *50S ribosomal protein L29 [Synechococcus sp. PCC 7002]*. The PM estimate of **7,858.347Da** induced by the respective aggregated path matches its theoretical mass upon truncation of the N-terminal methionine, being **7,859.344Da**, with a difference of **0.997Da**, which can be attributed to ±1Da errors often introduced in mass spectra at time of deconvolution. This protein must have appeared in the sample either as a carry-over protein from previous runs on *Synechococcus* on the same instrument or as a contaminant on the column.

Running time: on a 24-processor (Intel Xeon X5675, 3.07GHz) server with 190Gb RAM, *Twister* correctly retrieved the entire sequence of the alemtuzumab light chain in **47.85s**, being launched on the preprocessed mass spectra. The preprocessing stage required about 15min.

		top-down	bottom-up	
k		**4**	**4**	**3**
k-tags	aa sequences	921	14,127	4,839
	with offsets	4,005	57,549	100,212
T-Bruijn graph	vertices	2,794	30,361	49,214
	components	800	9,651	14,989

The 16 aggregated paths after validation and correction (red means correct):

LQSGNSQESVTEQDSKDSTYSLSSTLTLSKADYEKHKVYACEVTHQGLSS		
VFLFP	ERKLEVKTGQGFTR	SLTFTFDT
VSASLSSPSQTMQ	VSASLSSPSQTMQ	SLHQL
PVGTQLNNTNYLLL	KVDNAL	YQQK
PVTKS	CVVS	SGSGSGTD
KLQEP	QQYWN	NLDK

Conclusion

We have presented a fast and highly accurate method for *de novo* reconstruction of a protein sequence solely from MS data. Future research direction comprise automated determination of inner parameter values of the algorithm based on the input.

K.V., V.D., M.D., S.A., and P.A.P. were partly supported by Government of Russian Federation (grant 11.G34.31.0018). L.D. and M.V. were supported by the Netherlands Organization for Scientific Research (NWO), Zenith grant 93511034.

http://bioinf.spbau.ru/en/twister; vyatkina@spbau.ru