VarQuest: modification-tolerant identification of novel variants of peptidic antibiotics and other natural products

Alexey Gurevich[1], Alla Mikheenko[1], Alexander Shlemov[1],
Anton Korobeynikov[1], Hosein Mohimani[2] and Pavel Pevzner[1,2]

[1]Center for Algorithmic Biotechnology, Saint Petersburg State University, Russia
[2]Department of Computer Science and Engineering, University of California, San Diego, USA

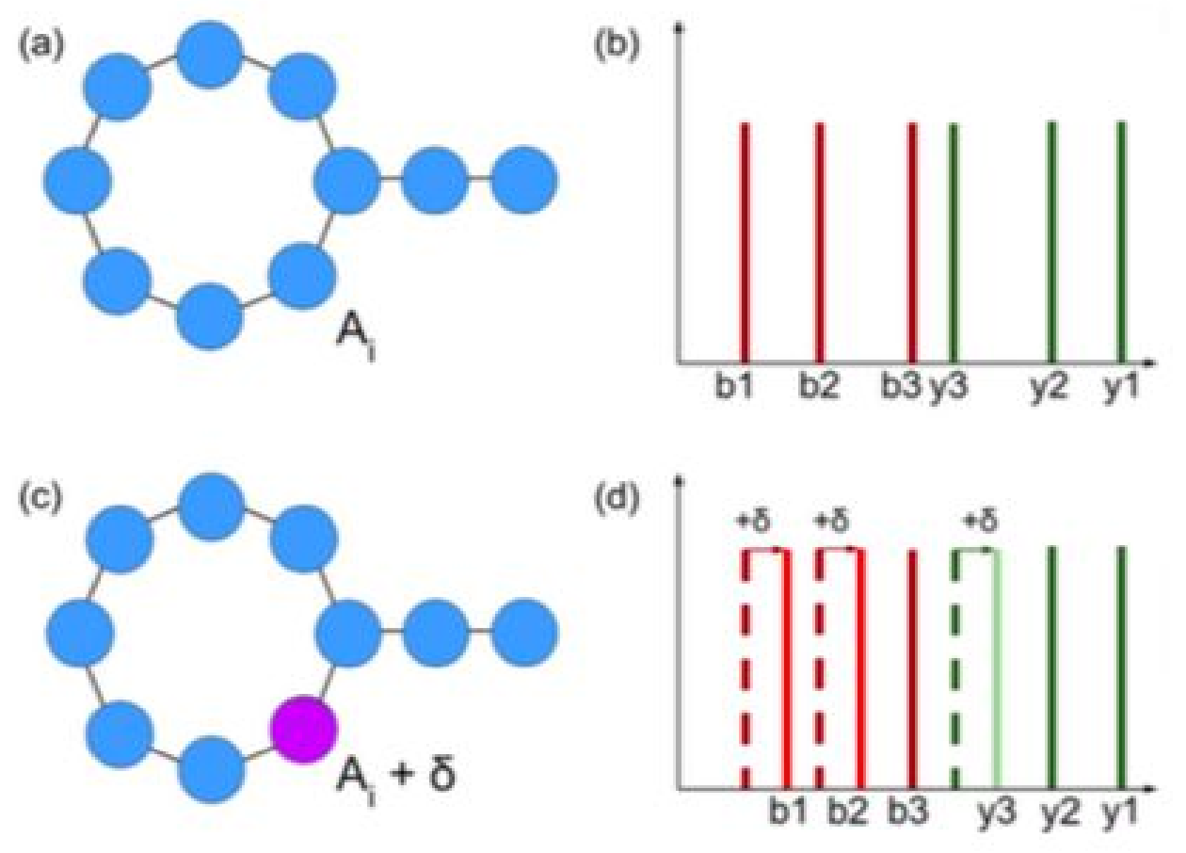
Motivation

Peptidic natural products (PNPs) include many antibiotics, anti-cancer agent, and other bioactive compounds. While billions of tandem mass spectra of natural products have been generated and deposited to Global Natural Products Social (GNPS) molecular network [1], discovery of novel PNPs and even variants of known PNPs from this gold mine of spectral data remains challenging. To address this problem, bioinformaticians develop dereplication techniques to identify known PNPs and their novel variants. However, since PNP databases are dominated by the most abundant representatives of PNP families, existing algorithms, focusing at dereplication of known PNPs, identify only a small fraction of spectra in the GNPS molecular network.

Scoring PNP-spectrum matches

- A *PNP graph* of a PNP *P* is defined as a graph with nodes corresponding to amino acids in *P* and edges corresponding to *generalized* peptide bonds.
- *TheoreticalSpectrum(P)* is defined as the set of masses (theoretical peaks) of all connected components of the PNP graph resulting from removal of two edges (a 2-cut) or a single edge (a bridge).
- *SPCScore(P, S)* is defined as the Shared Peak Count, the number of peaks shared between *TheoreticalSpectrum(P)* and experimental spectrum *S*.
- If *(A1,..., An)* is the list of amino acid masses in a PNP *P*, we define *P*=Variant(P, i, δ)* as *(A1,..., Ai + δ ,..., An)*, where *P* and *Variant(P, i, δ)* have the same topology and *Ai* + δ ≥ 0.
- *VariableScore(P, S)* is defined as *max(SPCScore(Variant(P, i, ω), S))*, where ω is *Mass*(*P*) − *Mass*(*S*) and *i* varies from 1 to |*P*| (the number of amino acids in the peptide *P*).

P and *P** correspondence



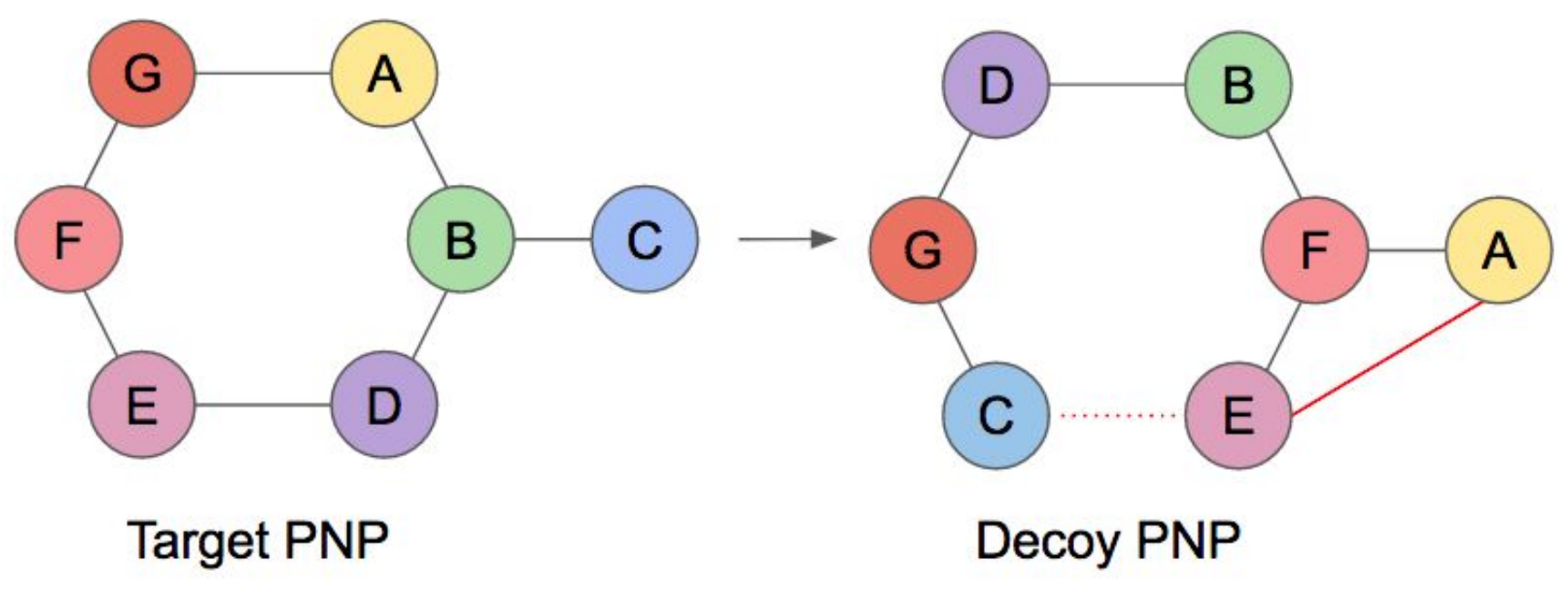
A PNP *P* (a) along with its theoretical spectrum (b) and *P**, a variant of *P* with a modification δ on one of its amino acids (c), along with its theoretical spectrum (d). For simplicity, only 6 peaks are shown.

TheoreticalSpectrum(*P*) and *TheoreticalSpectrum*(P^*) share approximately half of their peaks while the remaining peaks in *TheoreticalSpectrum*(P^*) are shifted by δ with respect to the corresponding peaks in *TheoreticalSpectrum*(*P*). If a spectrum *S* is produced by P^* and shares *N* peaks with *TheoreticalSpectrum*(*P*), we expect that *SPCScore*(*P*, *S*) ≈ *N/2* (not always true in practice).

FDR computation

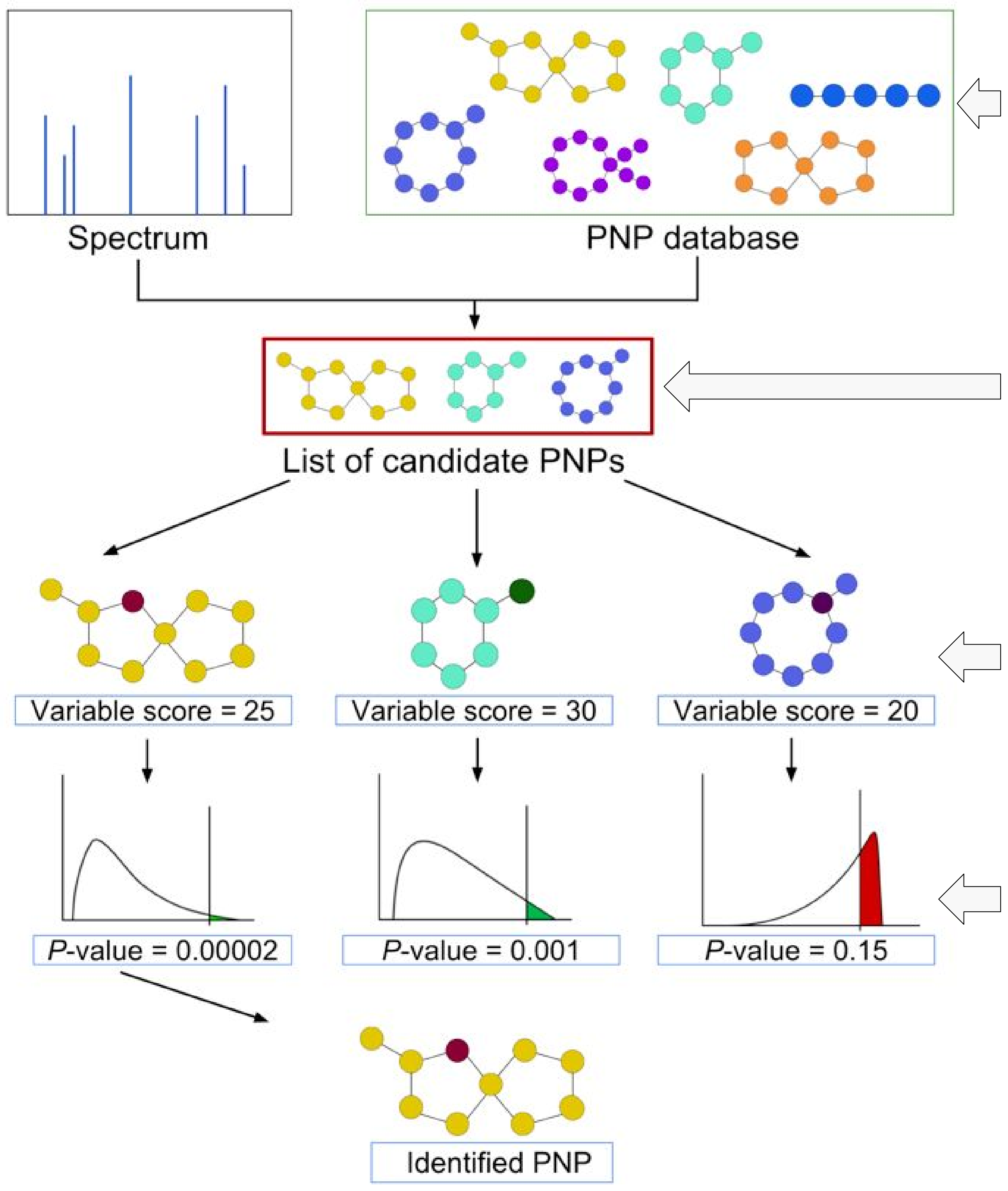
VarQuast estimates FDR using the target-decoy approach [2] and novel decoy database generation strategy to form decoy database *DecoyPeptides* from target database *Peptides*. We define PSM_τ (*Peptides*, *Spectra*) as the set of all PSMs found in *Peptides* (*DecoyPeptides*) and having *P*-values below τ, and compute FDR as:

$$FDR_\tau = \frac{|PSM_\tau(DecoyPeptides,\ Spectra)|}{|PSM_\tau(Peptides,\ Spectra)|}$$



Decoy generation is based on amino acid shuffling and random bond displacement

VarQuest pipeline



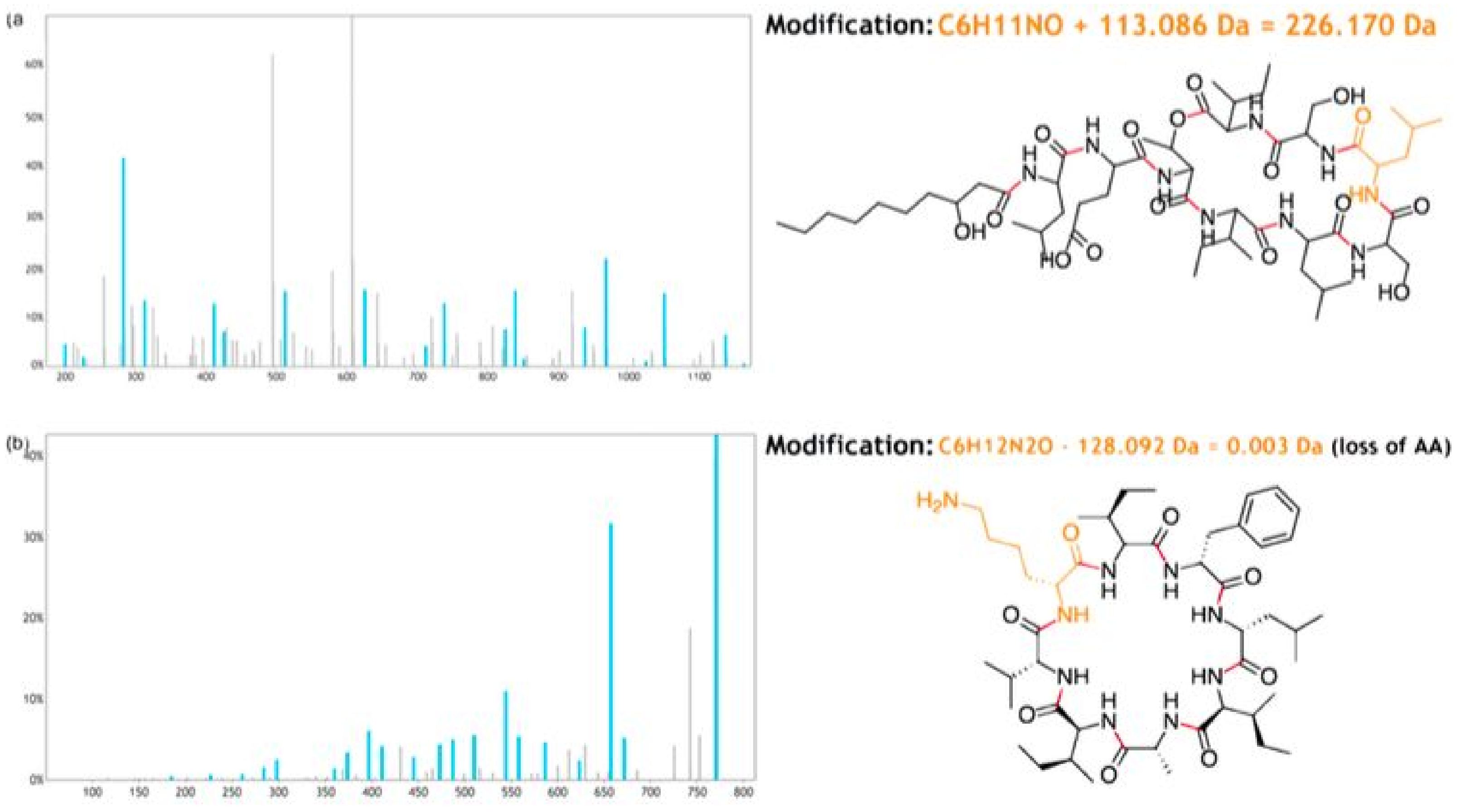
The *PNPdatabase* is constructed by combining all PNPs from AntiMarin, DNP, MIBiG and StreptomeDB databases.

For a given spectrum *S*, VarQuest includes PNP *P* in *CandidatePeptides(S)* if *P* satisfies:
Mass(S) − *MaxMod* < *Mass(P)* < *Mass(S)* + *MaxMod* (1)
and
SPCScore(*P*, *S*) ≥ η (2)
where (1) is checked based on a sorted list of PNP masses in the *PNPdatabase* and for fast calculation of (2) refer to [3].
Default values for *MaxMod* and η are 300 Da and 5.

For more details on the scoring procedure refer to [4].

P-value is computed using MS-DPR [5] and MCMC [6] algorithms.

Example identifications



Massetolide-1252. Reported by VarQuest as a novel variant of massetolide A (mass 1139.7 Da) with P-value 4.2 · 10^{-19} using a spectrum from *P. synxantha*. The suggested sequence is TISL+113SLI*EL (mass 1252.8 Da).
Confirmed by MS/MS and NMR analysis in the recent independent study [7].

Surugamide-769. Reported by VarQuest as a novel variant of surugamide B (mass 897.6 Da) with P-value 1.7 · 10^{-19} using a spectrum from *S. albus*. The suggested sequence is IAIVK^{-128}IFL (mass 769.5 Da).
Supported by genome mining results on *S. albus* using antiSMASH [8].

The entire GNPS dereplication

Search of 120 high quality datasets from the entire GNPS molecular network [1] against the *PNPdatabase* (5021 PNPs forming 1582 PNP families). *Standard* stands for standard identification with DEREPLICATOR [4]. P^{-10}, FDR_0 and FDR_5 stand for the number of identified PSMs, unique PNPs or PNP families with *P*-value below 10^{-10}, at 0% FDR and 5% FDR, respectively.

Method	# PSMs			# PNPs (families)		
	P^{-10}	FDR_0	FDR_5	P^{-10}	FDR_0	FDR_5
Standard	14757	7464	14746	420 (143)	279 (110)	420 (143)
VarQuest	379089	5661	179448	2673 (835)	675 (256)	2025 (648)

Availability

http://cab.spbu.ru/software/varquest

References

[1] Wang *et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking*. Nat. Biotechnol. (2016).
[2] Elias and Gygi. *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry*. Nat. Methods (2007).
[3] Shlemov et al. *Identification of novel peptidic antibiotics via large scale scoring of mass spectra against natural products databases*. Poster at RECOMB 2017 and ISMB 2017.
[4] Mohimani et al. *Dereplication of peptidic natural products through database search of mass spectra,* Nat. Chem. Biol. (2017).
[5] Mohimani et al. *A new approach to evaluating statistical significance of spectral identifications.* J. Proteome Res. (2013).
[6] Abramova and Korobeynikov. *Assessing the Significance of Peptide Spectrum Match Scores*. Proceedings of WABI 2017.
[7] Nguyen *et al. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides.* Nat. Microbiol. (2016).
[8] Medema *et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences*. Nucleic Acids Res. (2011).

Center for Algorithmic Biotechnology is supported by the Russian Science Foundation (grant #14-50-00069)