nature COMMUNICATIONS

ARTICLE

Received 2 Apr 2014 | Accepted 16 Sep 2014 | Published 31 Oct 2014

DOI: 10.1038/ncomms6277

MS-GF+ makes progress towards a universal database search tool for proteomics

Sangtae Kim[1,†] & Pavel A. Pevzner[1]

Mass spectrometry (MS) instruments and experimental protocols are rapidly advancing, but the software tools to analyse tandem mass spectra are lagging behind. We present a database search tool MS-GF+ that is sensitive (it identifies more peptides than most other database search tools) and universal (it works well for diverse types of spectra, different configurations of MS instruments and different experimental protocols). We benchmark MS-GF+ using diverse spectral data sets: (i) spectra of varying fragmentation methods; (ii) spectra of multiple enzyme digests; (iii) spectra of phosphorylated peptides; and (iv) spectra of peptides with unusual fragmentation propensities produced by a novel alpha-lytic protease. For all these data sets, MS-GF+ significantly increases the number of identified peptides compared with commonly used methods for peptide identifications. We emphasize that although MS-GF+ is not specifically designed for any particular experimental set-up, it improves on the performance of tools specifically designed for these applications (for example, specialized tools for phosphoproteomics).

[1] Department of Computer Science and Engineering, University of California San Diego, La Jolla, California 92093, USA. [†]Present address: Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. Correspondence and requests for materials should be addressed to P.A.P. (email: ppevzner@eng.ucsd.edu).

Mass spectrometry (MS) instruments and experimental protocols have greatly advanced over the last decade. New fragmentation technologies have emerged and high-precision mass spectrometers like Orbitrap have become widely available. Although trypsin remains a dominant protease in proteomics studies, digesting proteins with diverse proteases is becoming popular[1]. Empowered by these changes, MS researchers now have diverse choices with respect to the questions: 'what fragmentation method to use?', 'how accurate should be the measurements of the mass-to-charge (*m/z*) ratios?', 'what proteases to use?' and 'what post-translational modification (PTM) to focus on (for example, phosphorylation)?'. Depending on these choices, the resulting tandem mass (MS/MS) spectra vary in fragmentation propensities and precision. Therefore, unlike in the past when low-precision collision-induced dissociation (CID) spectra of tryptic peptides dominated the field, spectral data sets generated today are very diverse. Unfortunately, the popular MS/MS database search tools such as SEQUEST[2] and Mascot[3] have not kept pace with the increased diversity of the data. Although several new MS/MS database search engines were recently developed, including Andromeda[4], Morpheus[5], and MS Amanda[6], they have resulted in only minor improvements as compared with SEQUEST and Mascot.

Many efforts have been invested into making existing MS/MS search tools compatible with new types of data. For example, several pre- or post-processing strategies have been proposed[7,8], resulting in small improvement in the performance of database search tools. To further boost the performance, MS/MS database search tools are combined with statistical modelling tools such as PeptideProphet[9], Percolator[10] and IDPicker[11]. These tools do not find new peptide–spectrum matches (PSMs), but rather re-score PSMs reported by a database search tool using more complex scoring and output high-scoring PSMs. Although they often improve the performance of a database search tool, their performance is negatively affected when the database search tool fails to find correct PSMs[12]. Another downside of the pre- or post-processing strategies and statistical modelling tools is that, as they are often not integrated into database search tools, using them complicates the analysis of MS/MS spectra. Moreover, as different laboratories employ different combinations of tools (see Fig. 1), even for the same data, capabilities of analysing the data vary widely and results obtained in one laboratory are often difficult to reproduce in another laboratory[13].

In a recent review, Noble and MacCoss[14] pointed out that 'the field (of MS) is still missing a generic analysis platform that can be adapted automatically and in a principled manner, to handle spectra produced by any given fragmentation protocol'. Our MS-GF+ is a step towards achieving this goal, representing a universal database search tool that performs well for diverse types of spectral data sets. MS-GF+ works well (that is, identifies more peptides than other MS/MS tools that we tested) for spectra generated using diverse configurations of MS instruments and experimental protocols. However, the main contribution of this study is not the increase in the number of identifications for dozens of various fragmentation methods and experimental protocols but rather the fact that it represents the first truly universal MS/MS database search tool. We emphasize that MS-GF+ is not customized for specific spectral data sets but rather uses a robust probabilistic model that works well across all data sets.

MS-GF+ is universal because it automatically derives scoring parameters from thousands of PSMs without prior knowledge of the type of the spectra[12]. We represent various types of spectra as a graph where paths represent spectral types (Fig. 1). For each spectral type, MS-GF+ learns scoring parameters separately and scores a PSM using a different set of scoring parameters depending on the spectral type. MS-GF+ can train scoring parameters for any spectral type (including spectral types not specified in Fig. 1) or use pre-trained scoring parameters. It takes over the authority to train scoring parameters to the users and makes the training easy.

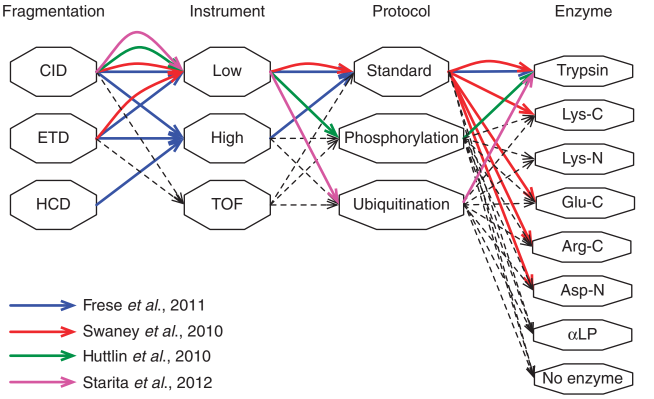

Figure 1 | Various spectral types. Spectral types are represented as paths in the graph representing possible choices of the fragment method (Fragmentation), the instrument measuring product ion *m/z* (Instrument), the protocol used to prepare a sample (Protocol) and the enzyme used to digest proteins (Enzyme). 'Low' in Instrument indicates low-resolution instruments (for example, linear ion trap), 'High' indicates high-resolution instruments (for example, Orbitrap) and 'TOF' indicates time-of-flight instruments. 'Phosphorylation' and 'Ubiquitination' in Protocol indicate that spectra are generated from phosphopeptides and ubiquitinated peptides, respectively. A path in the graph represents a spectral type. For example, the green path (CID, Low, Phosphorylation, Trypsin) represents low-precision CID spectra of trypsin digests generated from a sample enriched for phosphopeptides. The blue, red, green and magenta paths represent spectral types of the data sets used in recent studies by Frese *et al.*[20], Swaney *et al.*[1], Huttlin *et al.*[21] and Starita *et al.*[22], respectively. Different combinations of analysis tools were used for different studies. Frese *et al.* used an in-house tool for peak filtering, de-isotoping, and charge deconvolution, Mascot for database search, Percolator for re-scoring, and RockerBox[58] for peptide-level FDR control. Swaney *et al.*[1] used an in-house tool for peak filtering, OMSSA[27] for database search and an in-house tool for both peptide- and protein-level FDR control. Huttlin *et al.*[21] used an in-house tool for re-calibrating peak masses, SEQUEST for database search, an in-house tool for re-scoring and peptide- and protein-level FDR control. Starita *et al.*[22] used the Trans-Proteomics Pipeline[45] along with SEQUEST for database search. The same data sets were analysed by MS-GF+ without using any additional tool with scoring parameters trained separately for different spectral types.

The key advantage of MS-GF+ over existing approaches is its ability to compute rigorous *E*-values (using the generating function approach[15]) and thus to boost the number of peptide identifications. Although the generating function approach from ref. 15 worked well in a variety of studies[16–19], the question of applying it to modified peptides and to high-precision MS/MS spectra remains open. In this study we address these issues, thus making the generating function approach applicable to all types of spectra.

We demonstrate the performance of MS-GF+ using various previously studied data sets[1,20–23]: spectra of tryptic peptides generated using CID, higher-energy collisional dissociation (HCD) and electron transfer dissociation (ETD) in combination with either linear ion trap or Orbitrap readout, spectra of multiple enzyme digests, spectra of phosphopeptides and spectra or a novel protease alpha-lytic protease (αLP). For all these data sets, we show that MS-GF+ outperforms popular tools for peptide identification such as Mascot + Percolator.

Results

MS-GF+ scoring. Database search tools use a scoring function Score(P, S) to evaluate a PSM of a peptide P and a spectrum S, and further compute statistical significance of the resulting PSMs. In this study, we use E-values to evaluate statistical significance of individual PSMs (referred as spectral E-values) and the target-decoy approach to estimate false discovery rates (FDRs). See Gupta *et al.*[24] for the details of our probabilistic framework.

Let P_S be a peptide that generated S. A scoring function is adequate for S (with respect to a protein database *ProteinDB*) if the correct peptide attains the maximal score in the database, that is, $\max_{P \in ProteinDB}$ Score(P, S) = Score(P_S, S). A 'good' scoring function should satisfy the following three conditions. First, it should be adequate for the great majority of spectra. Second, the algorithm for PSM scoring should be fast. Third, the algorithm for computing statistical significance (for example, E-values) of PSMs should be fast and accurate.

MS-GF+ uses a very simple dot-product scoring Score(P, S) = $P^* \cdot S^*$ after converting peptide P and spectrum S into peptide vector P^* and spectral vector S^* (the spectral vector was called the prefix-residue-mass spectrum in the previous publications[12,25]). Conversion of a spectrum S into a spectral vector S^* uses a probabilistic model that ensures that the resulting dot-product scoring is adequate[26] (first condition). At the same time, it makes scoring and computing accurate E-values fast[15] (second and third condition). This simple 'dot-product' scoring model contrasts with many other database search[2,4,27,28] and re-scoring[9,10] tools, using sophisticated scoring functions that often make it difficult to satisfy the third condition.

MS-GF+ workflow. MS-GF+ takes a spectral data set *Spectra* and a protein database *ProteinDB* as an input and outputs a set of scored PSMs along with E-value estimates. It uses open source application programming interfaces jmzML[29], jmzReader[30] and jmzIdentML[31], and supports the HUPO Proteomics Standard Initiative standard file formats—mzML[32] and mzIdentML[33]. Owing to these developments, MS-GF+ has been already adopted in many proteomics pipelines and post-processing tools.

The workflow of MS-GF+ comprises the following four steps: generating spectral vectors, searching a protein database, computing E-values of PSMs and estimating FDRs. Below, we describe each step as well as how MS-GF+ takes advantage of high precision spectra.

Generating spectral vectors. A (non-modified) peptide is defined as a string over the alphabet $\mathcal{A}$ of 20 standard amino acids. Let $\mathcal{A}^+$ be an extended amino acid set containing both unmodified and modified amino acids. For an (unmodified) amino acid $a \in \mathcal{A}$, let Mod(a) $\subset \mathcal{A}^+$ be the set that contains a and all its modified amino acids. For example, if T (Thr) and T^* (phosphorylated Thr) are in $\mathcal{A}^+$, Mod(T) = {T, T^*}. Given a peptide $P = a_1 \ldots a_k$, define $PV = pv_1 \ldots pv_k$ as a variant of P if $pv_i \in$ Mod(a_i) for all i ($1 \leq i \leq k$).

MS-GF+ converts spectra into spectral vectors[12,25]. A spectral vector of a spectrum S is an M-dimensional vector with integer values, where M = PrecursorMass(S) is the nominal precursor mass of S. Here we consider nominal precursor masses, representing that the sum of nominal masses of amino acids of the peptide generated the spectrum. As in many cases, the precise nominal precursor mass is unknown (for example, MS instruments often choose second or third isotope peak instead of mono-isotope peak from MS1 spectrum), multiple spectral vectors are generated separately for each possible nominal precursor mass and the score of a peptide of mass M is computed from the spectral vector of precursor mass M.

The conversion from an experimental spectrum to a spectral vector proceeds as follows. A spectrum $S = \{(mz_1, \text{rank}_1), \ldots, (mz_l, rank_l)\}$ is represented as a set of ranked peaks where the ith highest intensity peak gets rank i (mz_j and rank$_j$ represent m/z and rank of jth peak, respectively). An ion type is represented as a triplet of integers charge, offset and sign, where sign represents whether the ion type is a prefix ion (sign = 1) or a suffix ion (sign = − 1). For example, singly charged b-ions and y-ions correspond to ion types (1, 1, 1) and (1, 19, − 1), respectively. Neutral losses and hydrogen transfers are also considered as ion types, for example, singly charged z·ions corresponds to (1, 3, − 1). Given an ion type $ion = (charge, offset, sign)$, one can turn a spectrum S into $S_{ion} = \{(mass_1, rs_1), \ldots, (mass_l, rs_l)\}$ using the following transformation:

$$mass_j = \begin{cases} [mz_j \cdot charge \cdot 0.9995] - offset & \text{if } sign = 1 \\ \text{PrecursorMass}(S) - ([mz_j \cdot charge \cdot 0.9995] - offset) & \text{if } sign = -1 \end{cases}$$

$$rs_j = \text{RankScore}(ion, rank_j),$$

where $[x]$ represents the closest integer to x and RankScore(ion, $rank$) is a pre-computed function that takes an ion type ion and an integer rank, and returns a probabilistic log-likelihood score defined in refs 12,26. It is noteworthy that 0.9995 is a rescaling constant for minimizing rounding errors (see Supplementary Table 1). In practice, RankScore(ion, $rank$) also accounts for the location of the observed peak and the precursor charge and mass of the spectrum, which are omitted here for simplification. Ion types contributing to scoring are selected from the training set as described in Kim *et al.*[12] Assume that $\mathcal{I}$ is a set of ion types that are selected. The spectral vector of S (denoted by $\mathbf{S} = (s_1, \ldots, s_M)$) is computed as follows:

$$s_i = \sum_{\text{ion} \in \mathcal{I}} \max(\{rs \mid (mass, rs) \in S_{ion} \text{ and } mass = i\} \cup \text{RankScore}(ion, \infty)),$$

where RankScore(ion, ∞) represents the score given when ion is missing.

We also define a peptide vector of a variant as follows. Let Mass(a) be the nominal mass of a (possibly modified) amino acid a. For example, Mass(T) = 101 and the mass of phosphorylated Thr is Mass(T^*) = 181. Given a variant $PV = pv_1 \ldots pv_k$, define the mass of PV as $\text{Mass}(PV) = \sum_{i=1}^{k} \text{Mass}(pv_i)$. Given a variant $PV = pv_1 \ldots pv_k$ of mass M, we define its peptide vector (denoted by **PV**) as a 0–1 vector $(m_1, \ldots, m_M)$ with $(n-1)$ 1 s, such that $m_i = 1$ if i equals to Mass(pv_1) + … + Mass(pv_j)($1 \leq j \leq k$).

The MS-GF+ score of a PSM (PV, S) is defined as $\text{MSGFScore}(PV, S) = \mathbf{PV} \cdot \mathbf{S} = \sum_{i=1}^{k} pv_i \cdot s_i$ if Mass(PV) = PrecursorMass(S) and $-\infty$ otherwise. The MS-GF+ score represents the log likelihood ratio described in ref. 26.

Searching a protein database. We define $ProteinDB^+$ as the set of all variants (with respect to an extended amino acid set $\mathcal{A}^+$) derived from *ProteinDB*. The goal of MS-GF+ database search is to solve the following problem: given a spectral data set *Spectra* and a protein database *ProteinDB*, for each spectrum $S \in Spectra$ find a variant $PV_{S,ProteinDB}$ such that

$$PV_{S,ProteinDB} = \underset{PV \in ProteinDB^+}{\text{argmax}} \text{MSGFScore}(PV, S).$$

In contrast to a traditional spectrum-based MS/MS database search approach that compares each spectrum against all peptides, MS-GF+ uses an alternative peptide-based approach that computes the suffix array to compare each peptide against all spectra with the same precursor mass. See Supplementary Note 1 for the details of MS-GF+ approach to the database search.

Computing *E*-values of PSMs. The scores of PSMs reported by existing MS/MS database search tools are often poorly correlated with their *E*-values[34]. It is important to rank PSMs based on their *E*-values, because such ranking (rather than ranking based on 'raw scores') often dramatically increases the number of identified spectra under a given FDR[15,35]. Many database search tools estimate an *E*-value of a PSM based on an approximation of a tail of the score distribution specific to the spectrum using peptides in the database[27,28]. As this approach is known to result in biased estimates of *E*-values[15], MS-GF+ adopted the generating function approach to rigorously compute *E*-values of PSMs using the score distribution of all peptides[15]. Our scoring model is essential here, because the generating function approach is easily applicable to the scoring functions that can be represented as a dot-product of vectors[24]. Adopting the generating function approach improves the accuracy of *E*-value estimates and increases the number of identified peptides as was recently confirmed by an independent work on applying it to the XCorr score in SEQUEST[35].

Given a spectrum S, a score threshold t, an extended set of amino acids $\mathcal{A}^+$ and a database size N, we define E-value (S, $\mathcal{A}^+$, t,N) as the expected number of variants PV (as defined by $\mathcal{A}^+$) with MSGFScore(PV,S)$\geq t$ in a random protein database of size N. To compute E-value(S, $\mathcal{A}^+$, t, N), we first compute *spectral E-value* E-value(S, $\mathcal{A}^+$, t), the expected number of variants PV with MSGFScore(PV, S)$\geq t$ given a single random peptide. A single random peptide models a random peptide starting at a fixed position in a random protein database.

We consider a set of all possible (unmodified) peptides of length k (where k is a large number) and select a random peptide uniformly from this set (that is, the probability of selecting a peptide is $\frac{1}{20^k}$). In practice, to reflect different frequencies of amino acids in a database (for example, Leu is usually more common than Trp), we define the probability of a peptide $P=a_1\ldots a_k$ as $\prod_{i=1}^{k}\text{Prob}(a_i)$, where Prob(a) is the frequency of amino acid a in a protein database. Note that this does not change the algorithm to compute the spectral E-values. We say that a peptide P produces a variant PV if PV is a variant of a prefix of P. For example, *PEPT** and *PEPTI* are produced by *PEPTIDE*. Given a spectrum S, let $\mathcal{PV}(t)$ be the set of all variants PV with MSGFScore(PV, S)$\geq t$. For every variant PV, there are $20^{k-|PV|}$ peptides of length k producing a variant PV ($|PV|$ stands for the number of amino acids in PV). Therefore, expected number of variants per random peptide with a score equal or better than t is

$$E-\text{value}(S,\mathcal{A}^+,t)=\sum_{PV\in\mathcal{PV}(t)}\frac{20^{k-|PV|}}{20^k}=\sum_{PV\in\mathcal{PV}(t)}20^{-|PV|}.$$

As a variant is a string over the alphabet $\mathcal{A}^+$, this expression can be computed using the generating function approach[15]. Given a spectrum S with $\mathbf{S}=s_1\ldots s_M$, consider a directed acyclic graph called an amino acid graph $G(V, E, \mathcal{A}^+)$ with $V=\{0, \ldots, M\}$ and $E=\{(i, j)|j-i\in\text{Mass}(a)\text{ for }a\in\mathcal{A}^+\}$, where the score of a vertex i is defined as s_i, the probability of an edge is defined as $\frac{1}{20}$, the score of a path is defined as the sum of scores of its vertices and the probability of a path is defined as the product of probabilities of its edges. A path in an amino acid graph represents a variant. Therefore, E-value(S, $\mathcal{A}^+$, t) equals to the sum of probabilities of all paths from 0 to M with scores equal or better than t, and can be computed using parametric dynamic programming[15,26,36].

Although spectral E-values are useful for evaluating statistical significance of individual PSMs (independently of the database), they need to be transformed into E-value(S, $\mathcal{A}^+$, t, N) to take into account the fact that the database search represents 'multiple testing' where multiple variants (arising from different database peptides) are scored against a spectrum[37]. E-values can be approximated as follows:

$$E\text{-value}(S,\mathcal{A}^+,t,N)\approx E\text{-value}(S,\mathcal{A}^+,t)\cdot N,$$

where N is the size of the database. It is noteworthy that as protein databases contain many repeated peptides, it is important to reflect the effective size of the database that is estimated as the number of unique peptides of certain length.

Estimating FDRs. MS-GF+ estimates FDRs using the target decoy approach[38,39]. See Supplementary Note 2 for details.

From low-precision to high-precision MS/MS spectra. Mass spectrometers are usually divided into high-precision (denoted by H) and low-precision (denoted by L) instruments. Depending on whether the precursor and product ions are measured with low or high precision, the spectra are divided into LL, LH, HL and HH spectra (LH spectra are hardly ever used in proteomics studies). Although it may appear that extending the generating function approach from LL (as defined in ref. 15) to HL and HH spectra is a simple matter of tuning parameters that control the error tolerance, the situation is more complex. Here we explain how MS-GF+ takes advantage of high-precision product ion peaks.

Let RMass(a) be the real mass of an amino acid a. For a variant $PV=pv_1\ldots pv_k$, let RMass(PV) $=\sum_{i=1}^{k}$ RMass(pv_i) and RPrecursorMass(S) be the real precursor mass of a spectrum S. We previously assumed that Mass(PV) and PrecursorMass(S) are integers and defined MSGFScore(PV,S)$=\mathbf{PV}\cdot\mathbf{S}$ if Mass(PV)$=$ PrecursorMass(S) and $-\infty$ otherwise. It is noteworthy that this condition, although appropriate for LL spectra, is weak for HL and HH spectra, because it may be satisfied even when the real mass RMass(PV) significantly deviates (for example, up to 0.5 Da) from RPrecursorMass(S). Let $a\overset{\Delta}{=}b$ represent the condition $|a-b|<\Delta$. To take advantage of accurate precursor masses in HL and HH spectra, the condition Mass(PV)$=$PrecursorMass(S) has to be redefined to RMass(PV) $\overset{\Delta}{=}$ RPrecursorMass(S), where Δ is the precursor mass tolerance. The database search problem with this modified definition of MSGFScore is now described by the following equation:

$$\begin{aligned}S_{PV,\text{Spectra}}&=\underset{S\in\text{Spectra}}{\text{argmax}}\,\text{MSGFScore}(PV,S)\\&=\underset{S\in\text{Spectra}_{\text{RMass}(PV)}}{\text{argmax}}\quad\text{MSGFScore}(PV,S),\end{aligned}\qquad(1)$$

where $Spectra_{\text{RMass}(PV)}$ represents the set of spectra $S\in Spectra$ satisfying RPrecursorMass(PV) $\overset{\Delta}{=}$ RMass(S).

The key part of the generating function approach is the assumption that amino acids have integer masses (otherwise the parametric dynamic programming is difficult to implement). However, rounding amino acid masses to integers introduces errors. These rounding errors reduce after rescaling by 0.9995, making them appropriate for LL and HL spectra. However, for HH spectra the rounding errors remain too large even after rescaling, prohibiting MS-GF+ from benefiting from precise product ion peaks. A larger rescaling constant could better accommodate the mass accuracy, for example, the rescaling constant 274.335215 allows one to model spectra with 2.5 p.p.m. accuracy[40]. However, as the time complexity of the generating function algorithm is proportional to the rescaling constant, this rescaling makes computing E-values prohibitively slow.

Here we present a new scoring algorithm taking advantage of the accurate product ion masses while not substantially increasing the running time of MS-GF+. In ref. 26, we introduced an abstract model (seemingly unrelated to MS) that described a probabilistic process of transforming a Boolean string (peptide vector) into another Boolean string (spectral vector). This model, although adequate for low-precision spectra, needs to be modified for high-precision spectra. Here we model a peptide as a Boolean string (as before) but model a spectrum as a directed acyclic graph (DAG) and further apply a transformation of a Boolean string into a DAG for scoring real PSMs (see Methods for details).

Our new idea behind the DAG modelling is as follows. Consider peaks at masses 100.01 and 157.4 that will be transformed into integer bins 100 and 157 in the Boolean string representation of the spectrum. After this transformation, we lose information about the exact difference between these two masses. However, in our new spectral DAG model this information is retained in edges of the spectral DAG and used in the scoring.

Data sets. Overall, we used 19 data sets (≈2.83 million spectra from human, yeast, mouse and *Schizosaccharomyces pombe*) reflecting the diversity of MS data, corresponding to 17 distinct spectral types shown in Fig. 1 (see Methods for details on the data sets). For all these data sets, we benchmarked MS-GF+ against popular tools for peptide identification such as Mascot+Percolator.

Comparison of MS-GF+ with Mascot+Percolator. We compared the numbers of identified PSMs at 1% FDR for MS-GF+ and Mascot+Percolator (that is, PSMs reported by Mascot and re-scored by Percolator). Mascot+Percolator (Mascot version 2.3.02 integrating Percolator) was used for the comparison, because it represents a popular choice for peptide identification. We also tested several other tools such as SEQUEST, InsPecT[25] and OMSSA but do not report their results because they identified significantly fewer PSMs as compared with Mascot+Percolator. See Supplementary Table S2 for database search parameters.

For all the 19 data sets, MS-GF+ identified significantly more PSMs compared with Mascot+Percolator (Fig. 2). Figure 3a shows the benchmarking results for the five human data sets generated with varying fragmentations and instruments[20]. Percolator greatly increased the number of identifications as compared with Mascot, but for all these data sets MS-GF+ identified significantly more PSMs (17–38%) than Mascot+Percolator (see Supplementary Fig. S1 for Venn diagrams of MS-GF+ and Mascot+Percolator identifications). We also compared the number of identifications reported by the original study[20], which also used Mascot+Percolator along with in-house pre- and post-processing tools. In this comparison, MS-GF+ also showed an improved performance (identifying 16–55% more PSMs).

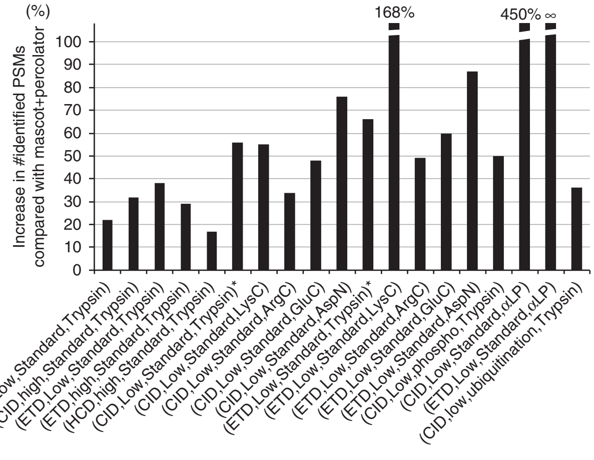

Figure 2 | Benchmarking MS-GF+ against Mascot+Percolator. Per cent increases in the number of identified PSMs for MS-GF+ compared with Mascot+Percolator for all 19 data sets. Each bar represents a spectral data set of a specified spectral type. For (CID, Low, Standard, Trypsin) and (ETD, Low, Standard, Trypsin), there are two corresponding data sets, one from human and the other from yeast. We distinguish them by adding '*' to the yeast data sets. For the (CID, Low, Phosphorylation, Trypsin) and (CID, Low, Ubiquitination, Trypsin) data sets, the number of phosphorylated and ubiquitinated PSMs were counted instead of the number of all identified PSMs. For the (ETD, Low, Standard, αLP) data set, Mascot+Percolator identified no PSM.

To figure out how each tool benefits from high-precision product ion peaks, for the three out of five human data sets representing HH spectra, we ran MS-GF+, Mascot+Percolator and Mascot using the parameters for HL spectra, that is, using 0.6 Da fragment mass tolerance for Mascot and Mascot+Percolator, and using the scoring model for low-precision spectra for MS-GF+. For every tool, the number of identifications was higher when the parameters for HH spectra were used, but the difference varied depending on the data set (Fig. 3b), and was negligible for ETD spectra.

Figure 3c shows the comparison for the ten yeast data sets generated with varying fragmentations (CID or ETD) and enzymes (Trypsin, LysC, ArgC, GluC or AspN)[1]. Again, for all these data sets, MS-GF+ identified significantly more PSMs (34–168%) than Mascot+Percolator (Fig. 3c). In ref. 1, using OMSSA (and in-house tools for pre- and post-processing), the authors reported the number of identified peptides at 1% peptide-level FDR that are matched to proteins identified at 1% protein-level FDR. We compared these numbers with the numbers of identified peptides at 1% peptide-level FDR using MS-GF+ (Fig. 3d). Note that this comparison is unfair because peptide identifications by MS-GF+ were not filtered out according to the protein that they are matched to. However, even after considering that, the results show that for most of the data sets MS-GF+ identified many more peptides than the original report.

To see whether our scoring model can capture the fragmentation propensities specific to phosphopeptides, we generated a scoring parameter set for (CID, Low, Phosphorylation, Trypsin). For the mouse data set corresponding to (CID, Low, Phosphorylation, Trypsin), we compared the numbers of identified PSMs for MS-GF+ with and without using the phosphorylation-specific parameter set, Mascot+Percolator and InsPecT equipped with a dedicated scoring model for (CID, Low, Phosphorylation, Trypsin)[41] (Supplementary Fig. S2a). Interestingly, without phosphorylation-specific scoring parameters, MS-GF+ outperformed both tools, identifying 37% and 44% more PSMs than Mascot+Percolator and InsPecT, respectively. With phosphorylation-specific parameters, MS-GF+ identified 9% more PSMs (and 12% more PSMs of phosphopeptides), confirming that our scoring model successfully captures phosphorylation-specific fragmentation propensities.

A similar result was obtained for a (CID, Low, Ubiquitination, Trypsin) data set (Supplementary Fig. S3). We emphasize that MS-GF+ does not 'know' anything about the peculiarities of the phosphorylation or ubiquitination, and simply trains the scoring parameters in exactly the same way it does for other spectral types. This ability to easily train modification-specific scoring parameters for any modification will greatly benefit MS researchers studying PTMs.

MS-GF+ for identifying peptides produced by a new protease. αLP is a new protease with cleavage specificities somewhat 'orthogonal' to trypsin[23]. MS-GF+ was applied to the study of αLP using two *S. pombe* data sets corresponding to (CID, Low, Standard, αLP) and (ETD, Low, Standard, αLP). We

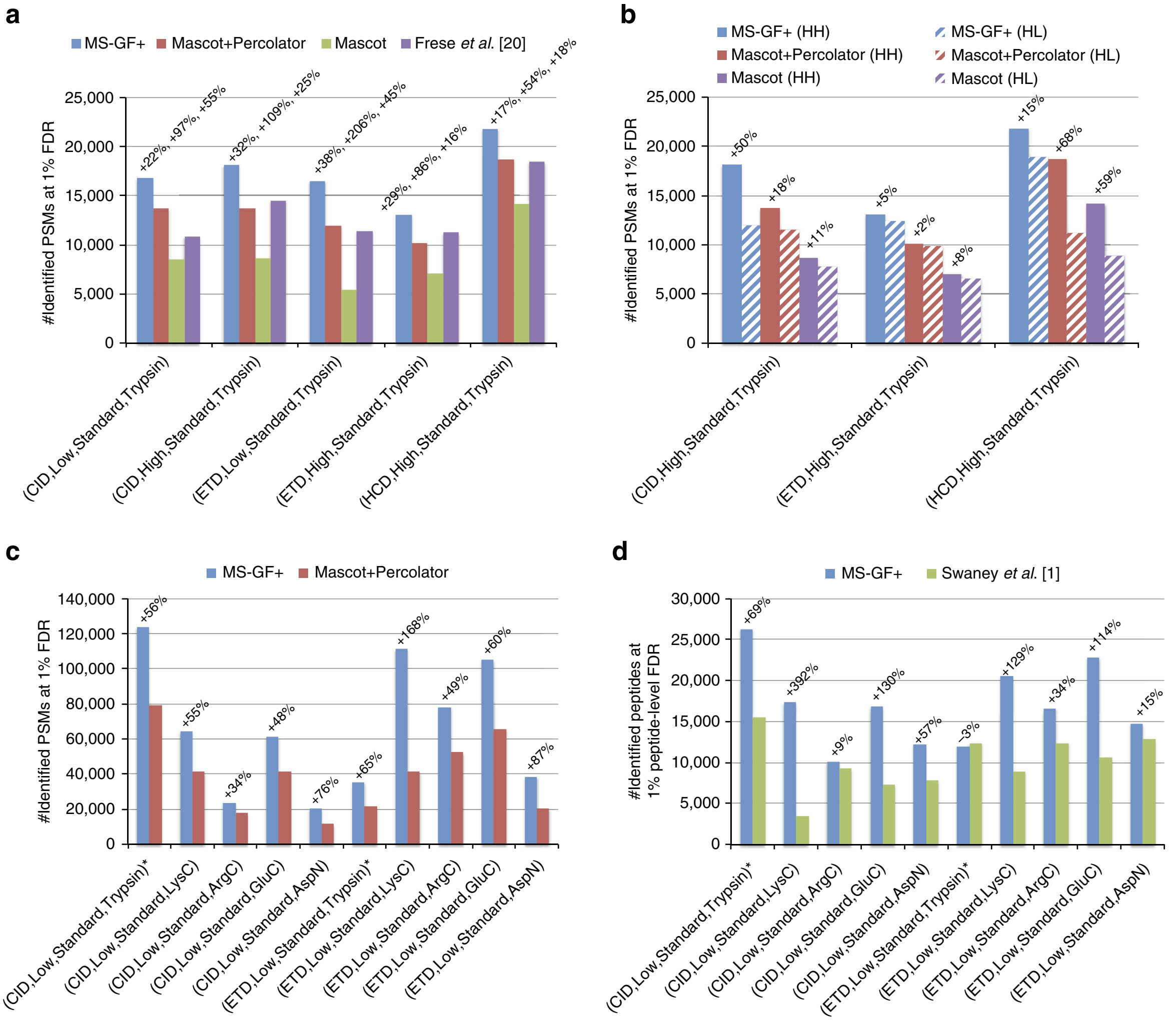

Figure 3 | Comparison of MS-GF+ and other tools for diverse spectral types. The numbers of identified PSMs (**a–c**) or peptides (**d**) at 1% FDR are shown. Numbers above bars represent the percentages of increase in the number of identifications for MS-GF+ compared with other tools. (**a**) Results for the human data sets with varying fragmentations and instruments. MS-GF+, Mascot+Percolator and Mascot results are shown along with the results in ref. 20. Percolator greatly increased the number of identifications as compared with Mascot, but MS-GF+ outperformed Mascot+Percolator for all the data sets. (**b**) Increase in the number of identifications due to the availability of high-precision product ion peaks. For the three human data sets representing HH spectra, MS-GF+, Mascot+Percolator and Mascot were run using search parameters for HL spectra. The results of these searches (denoted by HL) are compared with the numbers of identifications for the regular searches (denoted by HH). HH searches identified more PSMs than HL searches for every tool and every data set. The difference was larger for CID and HCD than ETD spectra. (**c**) Results for the yeast data sets with varying fragmentations and enzymes. MS-GF+ and Mascot+Percolator results are shown. MS-GF+ outperformed Mascot+Percolator for all these data sets. (**d**) Comparison of MS-GF+ and the results in ref. 1 that used OMSSA along with in-house post-processing tools for the yeast data sets. The numbers of (unique) peptides at the peptide-level 1% are shown. In ref. 1, only the number of identified peptides matched to proteins identified at 1% protein-level FDR was counted, while for MS-GF+ the number of identified peptides was counted regardless of their matched proteins.

ran Mascot+Percolator, OMSSA and MS-GF+ by specifying 'None' as an enzyme. As αLP produces peptides with different fragmentation propensities than tryptic peptides, Mascot+Percolator and OMSSA performed very poorly for this novel spectral type. In contrast, MS-GF+ identified 3,535 and 2,829 PSMs from the (CID, Low, Standard, αLP) and (ETD, Low, Standard, αLP) dataset using the scoring parameters for (CID, Low, Standard, Trypsin) and (ETD, Low, Standard, Trypsin), respectively (Supplementary Fig. S2b). The superior performance of MS-GF+ over Mascot+Percolator and OMSSA is because its scoring function is adequate for αLP peptides (correct peptide attains the maximal score) for a large portion of the spectra even when the search space is large (that is, no enzyme is specified). In fact, for the human data set corresponding to (ETD, Low, Standard, Trypsin), when no enzyme was specified and precursor mass tolerance 2.5 Da was used, MS-GF+ identified 10,937 PSMs, only 34% less as compared with the fully tryptic search with 7 p.p.m. precursor mass tolerance.

Using the identified PSMs by MS-GF+, we trained scoring parameters for (CID, Low, Standard, αLP) and (ETD, Low, Standard, αLP). When these αLP-specific scoring parameters were used, the number of identified PSMs further increased to 4,788 (+35%) and 3,313 (+17%) for (CID, Low, Standard, αLP) and (ETD, Low, Standard, αLP), respectively, showing the usefulness of MS-GF+ for studies of new proteases.

Thus, αLP represents a new alternative to trypsin, greatly increasing the PTM and protein sequence coverages, but generating spectra with unusual fragmentation propensities. We emphasize that the capabilities of αLP are not obvious when

Mascot + Percolator or another tool is used, because it fails to identify αLP peptides. The details on αLP protease have been discussed in a separate paper[23].

Running time of MS-GF +. We measured the running time of MS-GF + and Mascot + Percolator for LL, HL and HH spectra for various spectral types. For all the searches, MS-GF + and Mascot + Percolator showed similar running times (Supplementary Fig. S2c,d).

Discussion

Our analysis and recent independent studies[35,42–44] showed that for diverse types of spectral data sets, MS-GF + identifies more PSMs as compared with existing database search tools such as Mascot, X!Tandem, OMSSA, Crux, Comet and InsPecT, and statistical modelling tools such as Percolator. We emphasize that the generating function approach for accurately computing *E*-values significantly contributes to the improved performance of MS-GF +. For example, when *E*-values instead of MS-GF scores were used to cut off the results, the number of identified PSMs increased approximately by 70%, 50% and 20% for LL, HL and HH spectra, respectively.

Although we focused on demonstrating MS-GF + as a standalone tool, we emphasize that MS-GF + can be combined with various other proteomics analysis tools. As we have decided to release MS-GF + in 2012 well before this paper was prepared for a journal submission, MS-GF + has already been integrated into the following pipelines and statistical modelling tools: Trans-Proteomics Pipeline[45], Galaxy-P[46], ProteoSuite[47], IDPicker[11], SearchGUI[48], Scaffold[48], ProteoSAFe, Skyline[49] and Percolator[10,50]. Peptide identification tools that combine the results of multiple database search tools such as MSblender[51], Peptide-Shaker[52] and PepArML[53] also currently support MS-GF +. MS-GF + is freely available at http://proteomics.ucsd.edu.

Methods

Spectral DAG model. Given an extended alphabet $\mathcal{A}^+$, we first explain how to convert a spectrum S into a labelled DAG G. $G=(V,E)$ has a vertex set $V=\{0, \ldots, M=\mathrm{PrecursorMass}(S)\}$ and an edge set $E=\{(i,j) | j-i \in \mathrm{Mass}(a) \text{ for } a \in \mathcal{A}^+\}$. For simplicity, suppose that the set of ion types $\mathcal{I}=\{(1,0,1)\}$ (that is, only singly charged prefix ions with an offset zero contribute to the scoring). Given a constant δ called a fragment mass tolerance, two peaks of S with m/z x and y form a duo if $y-x$ is approximately equal to a mass of an amino acid, that is, $\mathrm{RMass}(a) \overset{\delta}{=} y-x$ for $a \in \mathcal{A}^+$. The vertex label s_i and the edge label $s_{i,j}$ of G are defined as follows: $s_i=1$ if there exists a peak of mass x satisfying $[0.9995 \cdot x]=i$ and $s_i=0$ otherwise; $s_{i,j}=1$, if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x]=i$ and $[0.9995 \cdot y]=j$, and $s_{i,j}=0$ otherwise (see Fig. 4 for an example).

Let $P=p_1 \ldots p_M$ be a Boolean string representing a peptide. Similar to the studies by Kim *et al.*[26] where a peptide string generates a spectrum string, we now assume that a peptide string generates a DAG. The probability of a peptide P generating a DAG G is defined as follows:

$$\mathrm{Prob}(G \mid P) = \prod_{i \in V} \mathrm{Prob}(s_i \mid p_i) \cdot \prod_{(i,j) \in E} \mathrm{Prob}(s_{i,j} \mid p_i, p_j),$$

where $\mathrm{Prob}(x|y)$ is a 2×2 matrix representing the probability of a peptide character y (0 or 1) generating a vertex label x and $\mathrm{Prob}(x|y,z)$ is a 2×4 matrix representing the probability of a pair of peptide characters y and z generating an edge label x (Table 1). In practice, $\beta_1 \approx \beta_2 \approx \beta_3$ (see Table 1b).

When applying this model for scoring a peptide P and a DAG G, we consider a test comparing two hypotheses: one assuming G is generated by P and the other assuming G is generated by an 'empty' string consisting of all zeros (denoted by O). The log-likelihood score of (P,G) (denoted Score(P,G)) is defined as follows (see Fig. 5 for an example):

$$\begin{aligned}
\mathrm{Score}(P,G) &= \log \frac{\mathrm{Prob}(G|P)}{\mathrm{Prob}(G|O)} \\
&= \log \frac{\prod_{i \in V} \mathrm{Prob}(s_i|p_i) \cdot \prod_{(i,j) \in E} \mathrm{Prob}\left(s_{i,j}|p_i,p_j\right)}{\prod_{i \in V} \mathrm{Prob}(s_i|0) \cdot \prod_{(i,j) \in E} \mathrm{Prob}\left(s_{i,j}|0,0\right)} \\
&= \sum_{i \in V} \log \frac{\mathrm{Prob}(s_i|p_i)}{\mathrm{Prob}(s_i|0)} + \sum_{(i,j) \in E} \log \frac{\mathrm{Prob}\left(s_{i,j}|p_i,p_j\right)}{\mathrm{Prob}\left(s_{i,j}|0,0\right)} \\
&\approx \underbrace{\sum_{i \in \{i | i \in V, p_i=1\}} \underbrace{\log \frac{\mathrm{Prob}(s_i|1)}{\mathrm{Prob}(s_i|0)}}_{\mathrm{VertexScore}(i)}}_{\text{vertex scoring}} + \underbrace{\sum_{(i,j) \in \left\{(i,j) \middle| (i,j) \in E, p_i=1, p_j=1\right\}} \underbrace{\log \frac{\mathrm{Prob}\left(s_{i,j}|1,1\right)}{\mathrm{Prob}\left(s_{i,j}|0,0\right)}}_{\mathrm{EdgeScore}(i,j)}}_{\text{edge scoring}}
\end{aligned}$$

Note that the last equation assumes that only the edges (i, j) with $p_i=p_j=1$ contribute to the edge scoring because $\beta_1 \approx \beta_2 \approx \beta_3$.

In practice, we generate multiple DAGs for a single spectrum, one for each ion $\in \mathcal{I}$. To generate an ion *DAG* for ion = (charge, offset, sign) with a real offset *r offset*, (for example, real offset of the singly charged b-ion is 1.008), we first convert $S=\{(mz_1, \mathrm{rank}_1), \ldots, (mz_l, \mathrm{rank}_l)\}$ into $S'=\{(mass_1, \mathrm{rank}_1), \ldots, (mass_l, \mathrm{rank}_l)\}$ using the following transformation:

$$\mathrm{mass}_j = \begin{cases} mz_j \cdot \mathrm{charge} - \mathrm{roffset} & \text{if sign} = 1 \\ \mathrm{RPrecursorMass}(S) - (mz_j \cdot \mathrm{charge} - \mathrm{roffset}) & \text{if sign} = -1 \end{cases}$$

Each peak of S representing ion corresponds to a peak of this converted spectrum S' representing an ion type (1, 0, 1). Therefore, the vertex and edge labels of the ion DAG for ion are defined as outlined before, but using S' instead of S (Fig. 4).

In reality, vertex and edge labels in the ion DAGs are integers rather than Boolean values. Given a converted spectrum S', we first remove all peaks (x, *rank*) if there exists another peak (x', *rank*$'$) where $[0.9995 \cdot x]=[0.9995 \cdot x']$ and $rank > rank'$. The vertex label s_i is defined as follows: $s_i = rank$ if there exists a peak (x, *rank*) satisfying $[0.9995 \cdot x]=i$ and $s_i=0$ otherwise. For an integer m, let AminoAcid(m) be the set of amino acids $a \in \mathcal{A}^+$ satisfying $\mathrm{Mass}(a)=m$ (for example, AminoAcid(128) = {Gln, Lys}). The edge label $s_{i,j}$ is defined as follows: $s_{i,j}=[100 \cdot \min_{a \in \mathrm{AminoAcid}(j-i)}(y-x-\mathrm{RMass}(a))]$ if there exists a duo of peaks with masses x and y such that $[0.9995 \cdot x]=i$ and $[0.9995 \cdot y]=j$, and $s_{i,j}=\infty$ otherwise. The constant 100 is multiplied to discretize the real-valued errors into bins of size 0.01 Da.

In this ion DAG representation, vertex labels encode the information on the intensities of individual peaks and the edge labels encode the information on the mass errors of pairs of peaks assuming they represent consecutive peaks of the same ion type. Note that edge labels take into account the spacing between peaks but do not take into account the peak intensities.

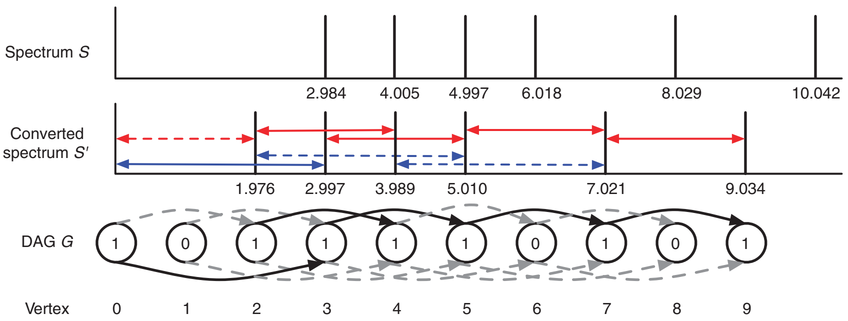

Figure 4 | Constructing a DAG in the case of two 'amino acids' with real masses 2.012 and 2.996. Assume that only singly charged b-ion with a real offset 1.008 contributes to the scoring. The spectrum S is converted into S' by shifting each peak by 1.008 to the left. Each arrowed line in S' represents a pair of peaks separated approximately by 2 Da (blue) or 3 Da (red) that form a duo (solid) or does not form a duo (dashed) for a fragment mass tolerance 0.01 Da. A DAG G is constructed from S'. The number in the vertex represents its label. The colour of the edge represents its label (0 for dashed grey and 1 for solid black).

Table 1 | Probability table for generating directed acyclic graphs

x \ y	0	1
0	θ	$1-\rho$
1	$1-\theta$	ρ

(a)

x \ y, z	0,0	0,1	1,0	1,1
0	β_1	β_2	β_3	$1-\alpha$
1	$1-\beta_1$	$1-\beta_2$	$1-\beta_3$	α

(b)

(**a**) Probability Prob($x|y$) of a peptide character y generating a vertex lavel x. (**b**) Probability Prob($x|y$, z) of peptide characters y and z generating an edge label x.

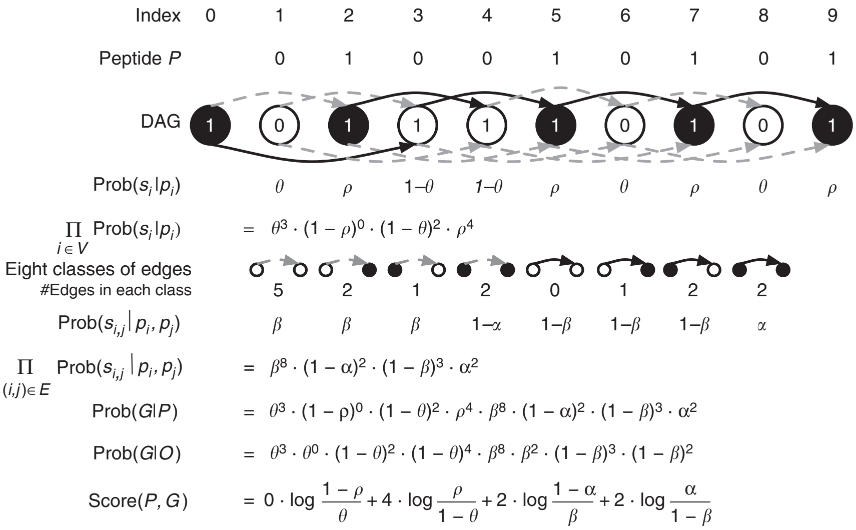
Index 0 1 2 3 4 5 6 7 8 9
Peptide P 0 1 0 0 1 0 1 0 1
DAG 1 0 1 1 1 1 0 1 0 1
Prob($s_i|p_i$) θ ρ $1-\theta$ $1-\theta$ ρ θ ρ θ ρ
$\prod_{i\in V}$ Prob($s_i|p_i$) $= \theta^3 \cdot (1-\rho)^0 \cdot (1-\theta)^2 \cdot \rho^4$
Eight classes of edges
#Edges in each class 5 2 1 2 0 1 2 2
Prob($s_{i,j}|p_i, p_j$) β β β $1-\alpha$ $1-\beta$ $1-\beta$ $1-\beta$ α
$\prod_{(i,j)\in E}$ Prob($s_{i,j}|p_i, p_j$) $= \beta^8 \cdot (1-\alpha)^2 \cdot (1-\beta)^3 \cdot \alpha^2$
Prob($G|P$) $= \theta^3 \cdot (1-\rho)^0 \cdot (1-\theta)^2 \cdot \rho^4 \cdot \beta^8 \cdot (1-\alpha)^2 \cdot (1-\beta)^3 \cdot \alpha^2$
Prob($G|O$) $= \theta^3 \cdot \theta^0 \cdot (1-\theta)^2 \cdot (1-\theta)^4 \cdot \beta^8 \cdot \beta^2 \cdot (1-\beta)^3 \cdot (1-\beta)^2$
Score(P, G) $= 0 \cdot \log\frac{1-\rho}{\theta} + 4 \cdot \log\frac{\rho}{1-\theta} + 2 \cdot \log\frac{1-\alpha}{\beta} + 2 \cdot \log\frac{\alpha}{1-\beta}$

Figure 5 | Illustration of the MS-GF+ DAG scoring. The peptide ABAA is converted into its Boolean string $P = 010010101$ and the spectrum S is converted into a labelled DAG G as described in the text. The number in the vertex represents its label. The colour of the edge represents its label (0 for grey and 1 for black). The vertex i is coloured depending on the peptide character i (white for 0 and black for 1). We also colour vertex 0 as black. The procedure to compute Score(P,G) is illustrated. All edges are partitioned into eight classes depending on $s_{i,j}$, p_i and p_j. For example, there are five edges with $s_{i,j} = p_i = p_j = 0$.

Supplementary Note 3 describes how to integrate information from various ion DAGs into a single spectral *DAG*.

Human data sets with varying fragmentations and instruments. Five human data sets corresponding to the spectral types (**CID**, **Low**, Standard, Trypsin), (**CID**, **High**, Standard, Trypsin), (**ETD**, **Low**, Standard, Trypsin), (**ETD**, **High**,Standard, Trypsin) and (**HCD**, **High**, Standard, Trypsin) contain 38,401, 33,586, 30,451, 25,734 and 37,810 spectra, respectively. These data sets are generated in the Heck Laboratory (Utrecht University). HEK293 whole-cell lysates were digested by trypsin and analysed by LTQ-Orbitrap Velos (Thermo Fisher Scientific, Bremen), using combinations of one of the three fragmentation modes CID, ETD and HCD, and either ion trap or Orbitrap readout for product ion *m/z*. The detailed experimental procedures are described in ref. 20.

Yeast data sets with varying enzymes. Ten yeast data sets corresponding to the spectral types (**CID**, Low, Standard, **Trypsin**), (**CID**, Low, Standard, **LysC**), (**CID**, Low, Standard, **ArgC**), (**CID**, Low, Standard, **GluC**), (**CID**, Low, Standard, **AspN**), (**ETD**, Low, Standard, **Trypsin**), (**ETD**, Low, Standard, **LysC**), (**ETD**, Low, Standard, **ArgC**), (**ETD**, Low, Standard, **GluC**) and (**ETD**, Low, Standard, **AspN**) contain 333,203, 278,336, 114,351, 81,669, 251,974, 72,463, 246,428, 204,860, 88,403 and 262,635 spectra, respectively. These data sets were generated in the Coon Laboratory (University of Wisconsin Madison). Yeast whole-cell lysates were digested separately, with either trypsin, LysC, ArgC, GluC or AspN, separated into 12 fractions via strong cation exchange chromatography and analysed in triplicate with an ETD-enabled LTQ-Orbitrap mass spectrometer, where peptide fragmentation was accomplished either with CID or ETD using the decision-tree acquisition mode[54]. We downloaded 180 (5 enzymes × 12 fractions × 3 replicates) spectrum files (Thermo RAW format) and converted each raw file into two mgf files, one containing CID and the other containing ETD spectra using 'msconvert' in ProteoWizard[55] with 'no filtering' option. The conversion was unsuccessful for 6 out of 180 files (5 from Arg-C and 1 from Glu-C digests). These six files were removed in the further analyses. The detailed experimental procedures are described in ref. 1.

Mouse data set of phosphopeptides. A mouse data set corresponding to the spectral type (CID, Low, **Phosphorylation**, Trypsin) contains 181,093 spectra. This data set was generated from the Gygi Laboratory (Harvard Medical School). Nine mouse organ proteins were digested with trypsin and the resulting peptides were fractionated via strong cation exchange. Phosphopeptides were enriched via immobilized metal affinity chromatography and analysed in duplicates via LC-MS/MS on an LTQ-Orbitrap mass spectrometer. Out of nine organ tissues analysed, we used the spectra generated from the brain tissue. The detailed experimental procedures are described in ref. 21.

***S. Pombe* data sets with αLP digest.** Two data sets corresponding to the spectral type (**CID**, Low, Standard, **αLP**) and (**ETD**, Low, Standard, **αLP**) contain 49,167 spectra each. These data sets were generated in the Komives Laboratory (University of California, San Diego). The detailed experimental procedures to generate these data sets are as follows. Wild-type *S. pombe* cells were lysed in 50 mM Tris–HCl pH 8.0, 150 mM NaCl, 5 mM EDTA, 10% glycerol, 50 mM NaF, 0.1 mM Na3VO4, 0.2% NP40 and stored at −80 °C. The debris was pelleted and then the supernatant was collected. The pellet was extracted according to ref. 56. Briefly, the pellet was resuspended in 200 μl of 0.1 M NaOH, 0.05 M ETDA, 2% SDS and 2% β-mercaptoethanol and incubated at 90 °C for 10 min. Acetic acid was added to 0.1 M and vortexed, followed by an additional incubation at 90 °C for 10 min before clarification by centrifugation and methanol/chloroform extraction. The pellet was resuspended in 100 mM Tris containing 0.1% sodium deoxycholate with TCEP at 5 mM. Free thiols were capped with *N*-ethylmaleimide. Excess reagent was removed by ultrafiltration with amicon-4 10 kDa centrifugal devices. The protein was then quantified and exchanged into 6 M guanidine for digestion overnight by αLP. The digests were quenched by the addition of formic acid to 1%, followed by desalting by sep-pak (Waters, Milford, MA). Peptides were then fractionated with electrostatic repulsion-hydrophilic interaction chromatography[57]. Fractions were assayed for protein concentration using a BCA assay and pooled into 18 fractions of equal protein concentration, evaporated to dryness and resuspended in 100 μl of 0.2% formic acid. Nano LC-MS/MS was performed with a LTQ XL mass spectrometer equipped with ETD. Ten microlitres of each fraction (≈1 μg) was injected onto a

12 cm × 75 μm I.D.C18 column prepared in house and eluted in 0.2% FA with a gradient of 5 to 40% Acetonitrile over 60 min followed by wash and re-equilibration totalling 90 min of MS data per run. The flow was split about 1:500 to a flow rate of about 250 nl min^{-1}. A survey scan was followed by data-dependent fragmentation of the four most abundant ions with both CID and ETD, with supplemental activation. The maximum MS/MS ion accumulation time was set to 100 ms. Fragmented precursors were dynamically excluded for 45 s with one repeat allowed.

Training scoring parameters. At the beginning of this study, we had five scoring parameter sets used in ref. 12 for the following five spectral types: (CID, Low, Standard, Trypsin), (CID, Low, Standard, LysN), (ETD, Low, Standard, Trypsin), (ETD, Low, Standard, LysN) and (ETD, Low, Standard, LysC). For this study, we constructed 20 new parameter sets using these 5 parameter sets as a starting point, using a newly developed programme called ScoringParamGen within the MS-GF+ package. To train scoring parameters for a new spectral type, MS-GF+ was run with an existing parameter set to identify PSMs at 1% FDR threshold, and using the identified PSMs as a training set a new parameter set was constructed. Supplementary Fig. S4 shows the scoring parameter sets contained in MS-GF+ and how they were constructed. We also tried to construct another generation of parameter sets using the existing parameter sets for the same spectral types, but this 'iterative training' hardly changed the number of identified PSMs.

For some data sets, the same data set was used for both training and testing of the performance, raising concerns about overfitting. However, as shown in ref. 12, MS-GF+ scoring parameter set characterizes a particular spectral type and is rather stable with respect to specific data sets. For example, for the human (CID, Low, Standard, Trypsin) data set, when the scoring parameter set trained from the same data set was used instead of the data set used in ref. 12, the number of identified PSMs hardly changed.

References

1. Swaney, D. L., Wenger, C. D. & Coon, J. J. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* **9,** 1323–1329 (2010).
2. Eng, J., McCormack, A. & Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
3. Perkins, D., Pappin, D., Creasy, D. & Cottrell, J. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
4. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the Maxquant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).
5. Wenger, C. D. & Coon, J. J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J. Proteome Res.* **12,** 1377–1386 (2013).
6. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **13,** 3679–3684 (2014).
7. Sweet, S. M. M. *et al.* Database search strategies for proteomic data sets generated by electron capture dissociation mass spectrometry. *J. Proteome Res.* **8,** 5475–5484 (2009).
8. Hsieh, E. J., Hoopmann, M. R., Maclean, B. & Maccoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **9,** 1138–1143 (2009).
9. Keller, A., Nesvizhskii, A., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392 (2002).
10. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & Maccoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4,** 923–925 (2007).
11. Ma, Z.-Q. *et al.* Idpicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **8,** 3872–3881 (2009).
12. Kim, S. *et al.* The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **9,** 2840–2852 (2010).
13. Yates, J. R. *et al.* Toward objective evaluation of proteomic algorithms. *Nat. Methods* **9,** 455–456 (2012).
14. Noble, W. S. & Maccoss, M. J. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.* **8,** e1002296 (2012).
15. Kim, S., Gupta, N. & Pevzner, P. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.* **7,** 3354–3363 (2008).
16. Zhou, J.-Y. *et al.* Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. *Anal. Chem.* **84,** 2862–2867 (2012).
17. Dresang, L. R. *et al.* Coupled transcriptome and proteome analysis of human lymphotropic tumor viruses: insights on the detection and discovery of viral genes. *BMC Genom* **12,** 625 (2011).
18. Wang, L. *et al.* Mapping N-linked glycosylation sites in the secretome and whole cells of aspergillus niger using hydrazide chemistry and mass spectrometry. *J. Proteome Res.* **11,** 143–156 (2012).
19. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337,** 1661–1665 (2012).
20. Frese, C. K. *et al.* Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an ltq-orbitrap velos. *J. Proteome Res.* **10,** 2377–2388 (2011).
21. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143,** 1174–1189 (2010).
22. Starita, L. M., Lo, R. S., Eng, J. K., von Haller, P. D. & Fields, S. Sites of ubiquitin attachment in *Saccharomyces cerevisiae*. *Proteomics* **12,** 236–240 (2012).
23. Meyer, J. G. *et al.* Expanding proteome coverage with orthogonal-specificity α-lytic proteases. *Mol. Cell. Proteomics* **13,** 823–835 (2014).
24. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22,** 1111–1120 (2011).
25. Tanner, S. *et al.* Inspect: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639 (2005).
26. Kim, S., Gupta, N., Bandeira, N. & Pevzner, P. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics* **8,** 53–69 (2009).
27. Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964 (2004).
28. Craig, R. & Beavis, R. C. Tandem: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).
29. Côté, R. G., Reisinger, F. & Martens, L. jmzMl, an open-source Java API for mzMl, the PSI standard for MS data. *Proteomics* **10,** 1332–1335 (2010).
30. Griss, J., Reisinger, F., Hermjakob, H. & Vizcano, J. A. jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics* **12,** 795–798 (2012).
31. Reisinger, F. *et al.* jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. *Proteomics* **12,** 790–794 (2012).
32. Martens, L. *et al.* mzMl-a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10,** R110.000133 (2010).
33. Jones, A. R. *et al.* The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11,** M111.014381 (2012).
34. Granholm, V., Noble, W. S. & Käll, L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *J. Proteome Res.* **10,** 2671–2678 (2011).
35. Howbert, J. J. & Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **13,** 2467–2479 (2014).
36. Jeong, K., Kim, S., Bandeira, N. & Pevzner, P. A. Gapped spectral dictionaries and their applications for database searches of tandem mass spectra. *Mol. Cell. Proteomics* **10,** 002220 (2011).
37. Noble, W. S. How does multiple testing correction work? *Nat. Biotechnol.* **27,** 1135–1137 (2009).
38. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214 (2007).
39. Jeong, K., Kim, S. & Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics* **13**(Suppl 16): S2 (2012).
40. Liu, X., Segar, M. W., Li, S. C. & Kim, S. Spectral probabilities of top-down tandem mass spectra. *BMC Genomics* **15**(Suppl 1): S9 (2014).
41. Payne, S. H. *et al.* Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis. *J. Proteome Res.* **7,** 3373–3381 (2008).
42. Stekhoven, D. J., Omasits, U., Quebatte, M., Dehio, C. & Ahrens, C. H. Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. *J. Proteom* **99,** 123–137 (2014).
43. Risk, B. A., Edwards, N. J. & Giddings, M. C. A peptide-spectrum scoring system based on ion alignment, intensity, and pair probabilities. *J. Proteome Res.* **12,** 4240–4247 (2013).
44. Lange, P. F., Huesgen, P. F., Nguyen, K. & Overall, C. M. Annotating N termini for the human proteome project: N termini and Nα-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *J. Proteome Res.* **13,** 2028–2044 (2014).
45. Deutsch, E. W. *et al.* A guided tour of the trans-proteomic pipeline. *Proteomics* **10,** 1150–1159 (2010).
46. Goecks, J., Nekrutenko, A., Taylor, J. & Team, G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
47. Gonzalez-Galarza, F. F. *et al.* A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. *OMICS* **16,** 431–442 (2012).

48. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11,** 996–999 (2011).
49. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968 (2010).
50. Granholm, V. *et al.* Fast and accurate database searches with MS-GF + Percolator. *J. Proteome Res.* **13,** 890–897 (2014).
51. Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A. I. & Marcotte, E. M. Msblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. *J. Proteome Res.* **10,** 2949–2958 (2011).
52. Kroksveen, A. C. *et al.* Cerebrospinal fluid proteome comparison between multiple sclerosis patients and controls. *Acta Neurol. Scand. Suppl.* 90–96 (2012).
53. Edwards, N. PepArML: A meta-search peptide identification platform for tandem mass spectra. *Curr. Protoc. Bioinformatics* **44,** 13.23.1–13.23.23 (2013).
54. Swaney, D. L., McAlister, G. C. & Coon, J. J. Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods* **5,** 959–964 (2008).
55. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536 (2008).
56. von der Haar, T. Optimized protein extraction for quantitative proteomics of yeasts. *PLoS ONE* **2,** e1078 (2007).
57. Hao, P. *et al.* Novel application of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) in shotgun proteomics: comprehensive profiling of rat kidney proteome. *J. Proteome Res.* **9,** 3520–3526 (2010).
58. van den Toorn, H. W. P. *et al.* RockerBox: analysis and filtering of massive proteomics search results. *J. Proteome Res.* **10,** 1420–1424 (2011).

Acknowledgements

We thank Jesse Meyer and Elizabeth Komives for providing the αLP data sets, and Nuno Bandeira, Matthew Monroe and Taejoon Kwon for providing feedback to improve MS-GF + . This work was supported by the US National Institutes of Health grant 1-P41-RR024851 from the National Center for Research Resources.

Author contributions

S.K. and P.P. designed the algorithms and the experiments, and wrote the manuscript. S.K. implemented the algorithms and performed the data analysis.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Kim, S. and Pevzner, P. A. MS-GF + makes progress towards a universal database search tool for proteomics. *Nat. Commun.* 5:5277 doi: 10.1038/ncomms6277 (2014).