The Paragon Algorithm, a Next Generation Search Engine That Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra*[S]

Ignat V. Shilov‡, Sean L. Seymour‡§, Alpesh A. Patel, Alex Loboda, Wilfred H. Tang, Sean P. Keating¶, Christie L. Hunter, Lydia M. Nuwaysir, and Daniel A. Schaeffer

The Paragon™ Algorithm, a novel database search engine for the identification of peptides from tandem mass spectrometry data, is presented. Sequence Temperature Values are computed using a sequence tag algorithm, allowing the degree of implication by an MS/MS spectrum of each region of a database to be determined on a continuum. Counter to conventional approaches, features such as modifications, substitutions, and cleavage events are modeled with probabilities rather than by discrete user-controlled settings to consider or not consider a feature. The use of feature probabilities in conjunction with Sequence Temperature Values allows for a very large increase in the effective search space with only a very small increase in the actual number of hypotheses that must be scored. The algorithm has a new kind of user interface that removes the user expertise requirement, presenting control settings in the language of the laboratory that are translated to optimal algorithmic settings. To validate this new algorithm, a comparison with Mascot is presented for a series of analogous searches to explore the relative impact of increasing search space probed with Mascot by relaxing the tryptic digestion conformance requirements from trypsin to semitrypsin to no enzyme and with the Paragon Algorithm using its Rapid mode and Thorough mode with and without tryptic specificity. Although they performed similarly for small search space, dramatic differences were observed in large search space. With the Paragon Algorithm, hundreds of biological and artifact modifications, all possible substitutions, and all levels of conformance to the expected digestion pattern can be searched in a single search step, yet the typical cost in search time is only 2–5 times that of conventional small search space. Despite this large increase in effective search space, there is no drastic loss of discrimination that typically accompanies the exploration of large search space. ***Molecular & Cellular Proteomics 6:1638–1655, 2007.***

From Applied Biosystems/MDS Sciex, Foster City, California 94404
Received, September 12, 2006, and in revised form, May 26, 2007
Published, MCP Papers in Press, May 27, 2007, DOI 10.1074/mcp.T600050-MCP200

This study presents a new software technology for the identification of peptides from tandem mass spectra called the Paragon™ Algorithm, hereafter referred to interchangeably as "Paragon." The most common application for this class of software tools is so-called "shotgun" or "bottom-up" proteomics experiments (1) where a protein mixture of any complexity is digested with a proteolytic enzyme or reagent, the peptides are analyzed by tandem mass spectrometry, and then software of this type is used to identify the peptides (2, 3) and, by inference, determine which proteins have been detected in the mixture (4).[1] Although it is currently much less common, this type of software can also be applied to the direct analysis of endogenous peptides that result from the natural proteolysis in an organism (5–10). The Paragon Algorithm and this study specifically focus on the peptide identification process. This search engine is part of a larger package called ProteinPilot™ Software, which uses the peptide identification approach described here and then automatically conducts protein inference analysis with the Pro Group™ Algorithm discussed elsewhere.[1–3]

Protein identification for the analysis of MS/MS fragmentation data in the bottom-up approach can be thought of as having four main stages: 1) preprocessing, 2) selection of peptide hypotheses, 3) scoring peptide hypotheses, and 4) protein inference. The preprocessing stage 1 can include conversion of raw data to simplified peak lists, averaging of spectra deemed sufficiently similar, filtering of spectra considered unlikely to yield a good identification, etc. Most tools fall into one of two main categories differing in how hypotheses are selected: *sequence* approaches use some *de novo*

[1] Seymour, S. L., Loboda, A., Tang, W. H., Nimkar, S., and Schaeffer, D. A. (2004) Poster presented at the 52nd ASMS Conference on Mass Spectrometry and Allied Topics, Nashville, TN (May 23–27, 2004).

[2] Seymour, S. L. (2005) PowerPoint presentation at the MCP Workshop: Criteria for Publication of Proteomic Data, Paris, France (May 12–13, 2005) (www.mcponline.org/misc/PariReport_PP.shtml).

[3] S. L. Seymour, A. Loboda, W. H. Tang, A. A. Patel, I. V. Shilov, and D. A. Schaeffer, manuscript in preparation.

estimation of sequence information from the observed MS/MS fragmentation (11–17), whereas *precursor* approaches rely on the precursor mass as the main filter (17–25). The goal of both approaches is to gain efficiency and discrimination by constraining the universe of all possible peptides and modifications to a much smaller search space that is tractable for scoring or manual inspection.

In sequence methods, an amino acid sequence(s) from manual or automated *de novo* sequencing of full or partial peptide sequences is used as an initial search space constraint. In the earliest example of this type of method by Mann and Wilm (11), a small section of sequence, referred to as a "sequence tag," would be manually interpreted and then provided to their algorithm along with the masses of the unsequenced regions flanking the sequence tag. They referred to all three pieces, preceding mass tag, sequence tag, and following mass tag, as a "peptide sequence tag." The database was subsequently scanned to find the matches to the three elements of the peptide sequence tag. In "error-tolerant" mode, all three elements of the peptide sequence tag are not required to match, allowing successful identification even in the presence of unsuspected modifications. At the same time, Pappin and co-workers were developing similar software (12), which now exists as the Mascot "Sequence query" search (26). This sequence-based approach has now been implemented in several forms, including MS-Seq in Protein Prospector (17). More recently, there have been sequence category approaches that use automatic *de novo* sequencing and attempt to call larger stretches of sequence particularly as a solution to the so-called "homology searching" problem where it is expected that the proteins from the species of interest are poorly represented in the database (27–29). Sequence tags have also been used to derive metrics of spectral quality and as part of the scoring step with precursor-type searches (30).

In the precursor category of algorithms, no MS/MS-derived sequence information is used, and peptide hypotheses are selected solely on the basis of conformance of the observed precursor mass to the mass of the theoretical peptide. The theoretical masses of all possible peptides are exhaustively enumerated given the database and search space constraints such as allowed modifications and digestion cleavage rules, and then *all* hypotheses that match the observed precursor mass within a prescribed tolerance are selected for scoring. Although this is a brute force approach, it is the dominant approach in current use, eclipsing approaches that use sequence tags. The two most common search engines, the "MS/MS ions" mode of the Mascot search engine (19) and the SEQUEST search engine (18), are of this type. The main reason for this is almost certainly the ease of automated analysis relative to sequence methods, which often require some manual sequencing.

Despite being less used, sequence tag algorithms should, in theory, be more powerful by increasing selectivity during hypothesis selection giving this type of algorithm the potential to be faster as well. However, in addition to being less practical for high throughput applications, sequence tags also come with a significant risk: an incorrect sequence tag call may exclude the right answer from consideration. Initially most tag-based methods relied on a single interpreted tag per spectrum where the assumption is made that the interpretation is correct. That is, the sequence information is used as a hard filter; portions of the database without this sequence are not considered. Newer tag-based approaches such as GutenTag (13) and InsPecT (14) have offered improvements by automatically determining sets of many smaller tags that are used to restrict to any sequences in the database that contain at least one of the tags.

Although precursor methods are broadly used, they do have significant limitations. Unlike sequence methods, the presence of a feature on a peptide that is not allowed in the search will prevent it from ever being identified. For example, if a peptide is N-terminally acetylated, but this feature is not allowed in the search, only wrong answers can be returned for a mass spectrum of this peptide. It might seem that the solution is simply to allow for a large number of variations in the search. This is not feasible, however, because it would bring with it a combinatorial explosion in additional wrong answers that would also need to be scored, yielding unacceptable search times and poor discrimination in scoring. In current practice, the upper limit of what is tractable with precursor-type search engines is around 6–10 modifications. Partly because of the challenges of large search space, current analyses typically only identify a fraction of the total MS/MS spectra acquired, roughly 5–20% for low resolution ion trap type instruments (3, 31) and 15–70% for quadrupole time-of-flight instruments (24, 32). In some cases, there may be 2–3-fold more spectra with sufficient fragmentation quality that go unidentified because of unexpected cleavages, incorrect monoisotopic peak assignments, incorrect charge state determinations, modifications and substitutions not considered, etc. Although the frequency of any single feature might be relatively small, collectively allowance for many less frequent features can account for a significant number of additional spectra, and thus it is desirable to find ways to improve exploration of large search space.

The Paragon Algorithm presents a new approach to protein identification. In contrast to recent advances in peptide identification, the algorithm relies on three key innovations that have nothing to do with the scoring stage. Our efforts have focused on the hypotheses selection stage, driven by the belief that there is greater potential for improvement from advances in determining *what to score*, not *how to score it*. First, the likely relevance of each sequence segment of a database to the MS/MS spectrum is quantified on a continuum using many weighted *de novo* sequence tags to compute

a Sequence Temperature Value (STV).[4] Second, feature probabilities are formally used to model the frequencies of peptide features such as modifications, digestion events, and substitutions, allowing the estimation of a net probability of any peptide hypothesis. The use of feature probabilities has also allowed a great reduction in the algorithmic complexity of the user interface through the implementation of a translation layer between what the user describes and what the engine understands. Third, an overall threshold is applied to the net effect of STV and feature probabilities, yielding a highly selective triage of which peptide hypotheses are worth scoring. The assessment of both tag evidence and feature probabilities on a continuum allows the efficient balancing of scoring effort to be commensurate with the likelihood that a candidate is worth scoring. Sequence regions more likely to be related to the correct answer for a spectrum are "searched more extensively" in the sense that peptide hypotheses with lower combined feature probabilities will be scored, whereas weakly implicated sequence segments are "searched less," only scoring precursor matches for peptides that have highly probable features.

The Paragon Algorithm offers significant advances in performance in searching very large search space and removes much of the informatics expertise barrier to doing quality protein identification by tandem mass spectrometry while maintaining the automation of conventional precursor-type search engines. The focus of this study was the fundamental description and validation of this new technology.

EXPERIMENTAL PROCEDURES

Sample Preparation—A mixture of proteins was assembled from 20 proteins purchased separately from Sigma-Aldrich and mixed at varied stoichiometries, several proteins at relative concentrations of 100, 20, 10, and 1, to cover 2 orders of magnitude of concentration. Protein mixture (1 mg/ml) in 50 mM NH_4HCO_3, 0.05% SDS was reduced in 0.4 mM tris(2-carboxyethyl)phosphine for 60 min at 60 °C. The cysteines were then alkylated with 1 mM iodoacetamide (Sigma) for 30 min in the dark at 20 °C. Porcine trypsin (Promega, San Luis Obispo, CA) with 2 mM $CaCl_2$ was added for a final enzyme to protein ratio of 1:25. The digest was conducted at 37 °C for 16 h and then desalted on a Poros R2 20 column. An aliquot was dried and submitted for amino acid analysis. The true number of detectable proteins in the sample was far greater than the nominal 20 that were added to the mixture (due to contaminant proteins present in the purchased stock) based on prior exhaustive analysis of this sample using multiple mass spectrometry techniques and careful resolution of the number of detectable isoforms for each protein. The true number of detectable protein forms in the sample was estimated to be ~130, and the true dynamic range of concentrations is likely to be over 3 orders of magnitude due to the additional contaminant proteins detectable in the purchased stocks.

Mass Spectrometry—The resulting peptide mixture was separated by reverse phase chromatography (Tempo™ nano-LC system, Applied Biosystems) using a 75-μm-inner diameter × 150-mm PepMap C_{18} column (Dionex) and a 30-min linear gradient from 5 to 30% acetonitrile in 0.1% formic acid with a total flow rate of 300 nl/min. The eluting peptides were ionized by electrospray ionization and analyzed by a QSTAR® Elite QqTOF system (Applied Biosystems/MDS Sciex). Peptide MS/MS spectra were acquired in an information-dependent manner utilizing the Analyst QS software 2.0 acquisition features (Smart Exit, rolling collision energy, and dynamic exclusion). The raw data file is included in the supplemental data.

Peak List Creation—Reduction of raw data in the *.wiff format to searchable MS/MS peak lists was conducted without any merging of putatively like spectra. No restriction of mass range for precursors was applied beyond the constraints used during acquisition. Spectra containing less than three fragment peaks were not searched. For the file examined in this study, no spectra were rejected. Peak lists are created automatically at the beginning of a search in ProteinPilot Software using this protocol. Mascot Generic Format peak list files (.mgf) generated from the raw data file in this study using both the 1.0 and 2.0 versions of ProteinPilot Software have been included in the supplemental data.

Mascot Searches—Mascot searches were performed from ProteinPilot Software version 1.0 to assure that exactly the same peak list was searched by both Mascot and the Paragon Algorithm. The Mascot server was version 2.1 and was run on a Dell Precision 340 computer with a Pentium IV 2.4-Hz processor, 1.0 gigabyte of RAM, and Windows XP SP2.

Paragon Searches—All Paragon searches were run using ProteinPilot Software version 1.0 on a Dell Latitude D810 laptop computer with a Pentium M 1.86-GHz processor, 2.0 gigabytes of RAM, and Windows XP SP2. To allow better comparison with Mascot and to avoid the issue of modification identification, custom modification sets that were depleted with respect to the normal operation of the software were created and used to more closely equal Mascot search space. Repetition of several of the same Paragon searches on the desktop computer used to run Mascot searches found that the two hardware configurations were fairly equivalent. Small search space Paragon searches ran 15% faster on the desktop configuration, whereas large search space searches ran about 17% slower on the desktop. These differences were relatively small, and the point of emphasis in the results is on the relative trends, not absolute speed measurement.

Annotation of Spectra for Performance Evaluation—An annotation was created for the reference file where the correct sequence was explicitly determined for a subset of the spectra in the whole file. The orthogonal nature of the protein information was leveraged to avoid bias toward either search engine while still allowing advantages to be detected. That is, a consensus set of confident proteins was determined from Mascot and Paragon-Pro Group analyses, and then only peptide IDs to these very confident proteins were included in the annotation. This approach allowed a natural distribution of fragmentation qualities to be included in the annotation, which thus contains a realistic distribution of low confidence to high confidence peptides. The goal was not to annotate every spectrum in the file, nor was it a goal to precisely determine the exact modification location; the aim was to identify only the correct sequence for each spectrum, accepting that this method is not perfect.

To accomplish this, protein identification analyses were conducted with both Mascot 2.1 and Paragon-Pro Group with the same search types later used for comparison, and the best peptide answers for each spectrum according to the best set of proteins were manually aligned for all searches. The only difference between the Paragon searches run for annotation and the searches used for comparison was that the normal set of 35 workup modifications was used for searches for annotation rather than the depleted sets. This yielded 1228 of the total 1987 spectra (62%) with an answer in at least one of

[4] The abbreviations used are: STV, Sequence Temperature Value; ROC, receiver operating characteristic; ID, identification; SS, Search Space; CDS, Celera Discovery System.

the searches. Note that these were not necessarily the top ranked peptides for each spectrum. Each spectrum was manually validated for the presence of an answer with sufficient orthogonal evidence to be included in the annotation without risk of bias toward the engine that produced it if it was found by only one of the engines. Then the intended grading protocol was run on all of these searches, and all cases where either engine reported a high confidence answer that was graded as incorrect were inspected manually. There were few of these cases, and the majority of them were due to Lys/Gln differences or absence of one of these forms in the searched database. Because of this, we decided to allow Lys/Gln difference during grading.

Each peptide answer in the annotation had to be associated with a multi-hit protein or have a clear consensus peptide identification between the two engines, and the vast majority had both conditions. This reduced the set of 1228 spectra down to 902 of which an additional 12 spectra were excluded because spectra with ambiguous charge state assignments were not handled properly in submitting peak lists to Mascot in the first version of the software. Ultimately this left 890 spectra that were included in the annotation of which 708 (80% of 890) had correct answers that were sequences found by both search engines, not necessarily in the same spectrum. The other 182 (20%) of the annotated spectra were sequences from Paragon only, but they were from proteins clearly found by Mascot and had at least 50% confidence in one of the Paragon Algorithm searches. Because the full workup modification sets were used for annotation but not the main series of searches in this study, 96 of these 182 were out of search space for both search engines because the right answers had modifications that were not allowed. Most of these additional modifications were from minor side reactions of iodoacetamide such as modification of peptide N termini and reaction with methionine followed by dethiomethylation. Of the spectra in the annotation, 90% were associated with the top 32 proteins in the Paragon Thorough search of the CDS Combined database, meaning the vast majority of annotated spectra were connected to extremely solid protein identifications. The peptide set generally had few missed cleavages with 91% having none, 8% having one, and 1% having more than one missed cleavage. Because the file was a relatively deep characterization in terms of the number of spectra per proteins detectable in the sample and because the annotation set is enriched for peptides from multi-hit proteins, the frequency of cleavages at sites other than tryptic specificity was moderately high with 70% fully tryptic, 29% semitryptic, and less than 1% fully non-tryptic. The annotation and additional statistics are included in the supplemental data.

Grading Searches against the Annotation—All search results were graded against the annotation for only the subset of 890 spectra for which the right answers were known. The grading protocol compared the peptide sequence of each answer against the known correct sequence(s) for the spectrum allowing for bidirectional Ile/Leu and Lys/Gln substitution and unidirectional Asn → Asp and Gln → Glu to allow for equivalence via deamidation. It was determined that more than one correct sequence should be allowed for 12 spectra (1.3% of 890) because manual inspection of the spectra showed they lacked fragmentation information that could favor a single answer. Virtually all of these cases had a pair of or several shuffled residues. The exact modification state, including name and location, was not considered as part of the grading procedure, consistent with the effort to remove the issue of modification finding throughout this study.

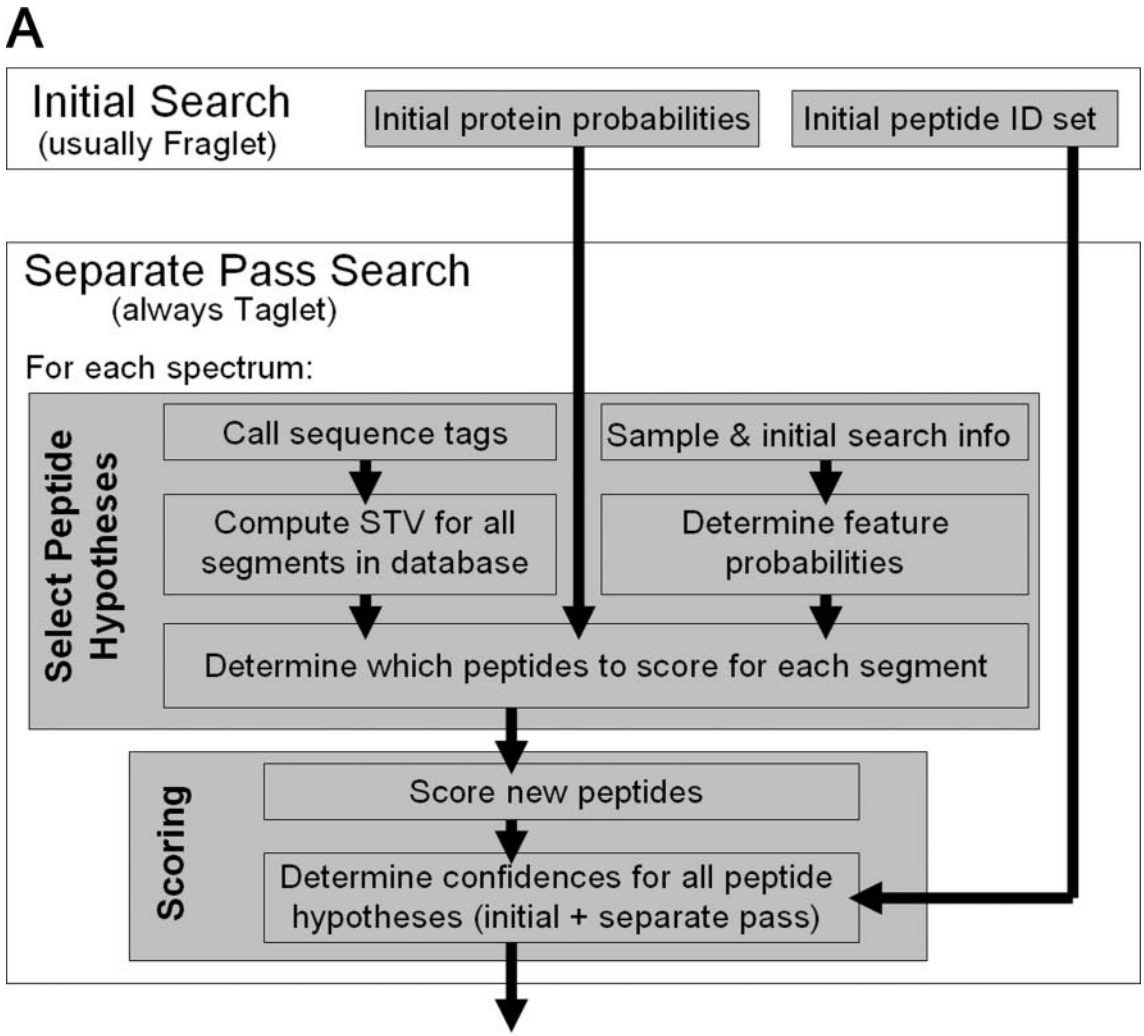
Receiver Operating Characteristic (ROC) Analysis—ROC data were generated for a search by taking all first reported peptide answers for each spectrum, sorting the list by the peptide discriminating variable of the search engine, and tallying the cumulative sum of correct and incorrect first answers according to grading against the annotation, moving from highest to lowest confidence. The discriminating variable for the Paragon Algorithm is the peptide Confidence value, which is a 0–99.0 scaled real number. The peptide E-value was used as the discriminating variable for Mascot. This was chosen over the ion score because it takes advantage of spectrum-specific significance thresholds. Note that only the first answer was considered; other degenerate top ranked answers were not considered. This is necessary because engines may vary in their granularity of binning in ranking answers.

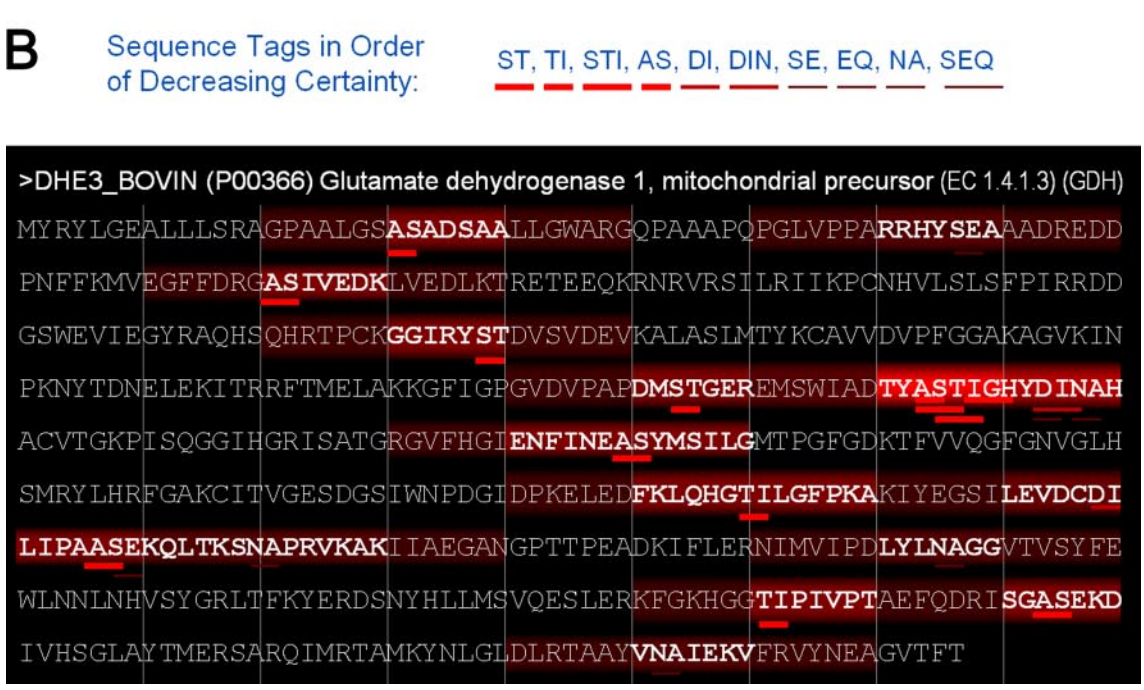
RESULTS

Paragon Algorithm Search Components—Fig. 1*A* presents a diagram of the elements of a Paragon search. The peptide identification algorithm has two core components that are invoked depending on the needs of the particular search. The first component, referred to as Fraglet, is essentially a standard precursor mass-filtered database search that can be run in isolation as the initial search in conjunction with the second component or not at all as in the case of no digest searching. The second component, referred to as Taglet, is the sequence tag component. This component is always the "separate pass" and is the initial search as well for no digest searches. In all cases, the coordination of which components are used is controlled automatically by the software based on user input. Unlike our previous Interrogator algorithm (24), both components were designed to work directly from the database without any need to first create an index file. Although the indexing made some types of searching faster, it was decided that the greater flexibility to support different digestions, species filters, and modifications was more important. A common scoring method is used in the two search components. A peptide hypothesis is scored with a *p* value giving an absolute measure of the chance a hypothesis might randomly match as many fragment ions to the observed spectrum, ignoring homology. This is generally done using only b and y ions. A percent confidence for a peptide is determined by taking into account the quality of all other matches derived for the same spectrum, how distinct these matches are from each other using a basic homology measure, and the probability of the various feature attributes for each peptide, *j*, as given by Equation 1.

$$\text{Confidence}_j = \frac{f(p\text{-value})_j(p_{\text{hypothesis}})_j}{\sum_{i=1}^{n}(f(p\text{-value})_i(p_{\text{hypothesis}})_i)} \qquad \text{(Eq. 1)}$$

The summation in the denominator includes only one member for each set of highly identical peptides. This allows a set of very similar high quality matches to all have high confidence (where generally only one among the ambiguous set is actually right) while it brings a beneficial competitive element that dilutes the confidences in cases with many dissimilar marginal matches. The probability of a peptide hypothesis, $p_{\text{hypothesis}}$,





is determined by information independent of MS/MS fragmentation,

$$p_{\text{hypothesis}} = \prod_{f=1}^{m}(p_f) \quad \text{(Eq. 2)}$$

where p_f are probability factors for various features of the peptide hypothesis such as modifications or lack of expected modifications, conformance of peptide termini to expected digestion patterns, and consistency of the observed precursor ion to the theoretical mass to charge ratio. For example, a tryptic peptide with the expected cysteine alkylation modification would have a much higher hypothesis probability than a peptide with neither end conforming to tryptic digestion and missing an expected modification. We estimate the p_f factors by empirically measuring the fraction of occurrences of a feature. For example, the probability of cleavage between lysine and proline could be estimated as follows.

$$p_{\text{cleavage (K-P)}} = \frac{\upsilon_{\text{cleaved K-P}}}{\upsilon_{\text{cleaved K-P}} + \upsilon_{\text{uncleaved K-P}}} \quad \text{(Eq. 3)}$$

Clearly the frequencies of features will vary even among data sets that are putatively treated and acquired the same way. We have found that for the various ways feature probabilities are used by Paragon, estimating average values by looking at many samples of the same type captures enough of this variation, and more importantly, Paragon has proven to be quite robust such that rough estimates are sufficient.

Although higher precision is used internally, peptide confidences are never reported higher than 99.00% to place a limit on the impact any single peptide can have on protein identification. This ceiling is justified as a conservative error bound because the accuracy rate in any subset of spectra is almost never higher than this.

Taglet Search Component—For a given spectrum, a substantial number of sequence "taglets" two and three amino acids long are called. Each tag is rated on a quality scale to indicate how likely it is that the tag call is correct. In this case, the tags are generated by doing *de novo* sequencing and breaking the results into continuous or nearly continuous sequence sections, although there are clearly other methods for automated derivation of sequence tags that could be used. Similarly the approach could be used with variable or longer length tags than have been used here. Modified amino acids are used in tag calling where the set of allowed modifications is determined automatically by applying a threshold to the estimated modification feature probabilities. Each called tag is then matched against the database to find all locations where the sequence occurs. The sequences in the database are divided into segments seven residues in length. Longer, shorter, or even variable length segments could be used instead. The degree to which a segment is implicated by the set of tags is evaluated as the STV for that segment,

$$\text{STV}_i = T_i + cT_{i+1} + cT_{i-1} \quad \text{(Eq. 4)}$$

where T_i is the net evidence or score from all taglets mapping to segment i calculated as

$$T_i = \sum_{j=1}^{n} t_j \quad \text{(Eq. 5)}$$

the sum of tag quality scores t_j for all n tags that map to segment i. Of course, there are many ways to determine a "net effect" of the evidence of many tags of differing qualities. The second and third terms in Equation 4 allow the neighboring segments of a segment in the database to influence the STV of the segment by including their T scores diminished by some fractional coefficient, c.

The calculation of STV for all sequence segments in the database allows them to be ranked, producing a full range of the degree to which each segment is implicated by the set of tags. Segments that are closer to the true sequence of the correct peptide for the spectrum should be ranked higher because more and higher quality tags hit these segments, whereas segments that are unlikely to be related to the correct peptide should be ranked lower because fewer and

Fig. 1. *A*, Paragon Algorithm diagram. The two major *boxes* depict the separate pass approach of the Paragon Algorithm where an initial search is done, typically a Fraglet search, and then all spectra are searched again in a separate pass, which is always a Taglet search. The *separate pass box* gives the steps in a Taglet search of one spectrum. A large set of short tags are called and then used to compute STVs for all segments in the database; all possible peptides are computed; feature probabilities for digestion events, modifications, and mass deltas are determined from the inputted sample information; and a decision is made to score or not to score each peptide hypothesis if its overall probability is greater than a threshold value. The overall probability is based on the STV of the segment, the features probabilities of the peptide, and prior probabilities of the protein from which the peptide is derived as estimated from the initial search. Then all scored peptide hypotheses for the spectrum from both the initial search and the separate pass are considered to assign confidences to all the scored hypotheses. *B*, a simplified example of the sequence temperature concept. A simplified list of sequence tags called from a spectrum are shown in order of decreasing confidence, and the mapping of these taglets to a single example protein sequence are indicated by *underlining* proportional to confidence. The protein sequence is divided into segments seven residues long as indicated by the *vertical lines*. The Sequence Temperature Value of each of these segments is indicated by the *degree* of *red glow*. The two segments with the highest STVs, TYASTIG and HYDINAH, have the most collective tag evidence and also benefit from a proximity effect as described in Equation 4. For these hot segments, even peptide hypotheses with very low probability features will be scored because the segments are very likely to be related to the true answer, whereas at the other limit of the spectrum, only extremely probable peptides will be scored for cold segments like MYRYLGE at the beginning of the protein.

poorer tags match these segments.

Peptide hypotheses are then generated from each segment using all allowed features regardless of probability, and then the overall probability for each hypothesis being the correct answer for the spectrum is calculated using the fundamental Paragon equation.

$$p_{\text{overall}} = p_{\text{segment}} p_{\text{hypothesis}} p_{\text{protein}} \quad \text{(Eq. 6)}$$

The probability that the segment used to generate a peptide hypothesis is associated with the correct answer, p_{segment}, is determined based on the STV ranking of the segment among all segments. The probability p_{protein} that the protein corresponding to the peptide hypothesis is detected in the sample as estimated by the initial search can also be factored into the decision to score or not to score a peptide.

By applying a threshold to the fundamental equation, Equation 6,

$$p_{\text{threshold}} < p_{\text{overall}} \quad \text{(Eq. 7)}$$

scoring can efficiently be limited to only those peptides that have an overall probability that assures a minimum level of believability while at the same time letting search space be very large for those segments very likely to contain the true sequence. This is *the* central innovation in the Paragon Algorithm. Sequence segments with very "hot" STVs are searched addressing very large search space such that peptide hypotheses containing lower probability features such as unlikely modifications and unexpected cleavages will be considered. Segments at the other limit of "cool" STVs are searched within very small search space such that only peptide hypotheses with the most probable features will be considered. These ideas are illustrated in Fig. 1*B*.

Note that because precursor mass delta is a factor, this approach is able to consider hypotheses that differ greatly from the expected mass. Robustness to inaccurate precursor mass information is one of the conventional advantages of sequence-based approaches over precursor-based methods, and this approach preserves that benefit. As long as the net effect of STV and other factors is favorable, large delta hypotheses can be considered in scoring. This allows identifications that would be lost by other approaches to be recovered, for example, in cases where multiple peptides pass the first mass analyzer.

Because the algorithm uses many tags and considers their qualities, identifications can also be recovered for some cases where the exact sequence is not in the database or the appropriate modifications were not considered. Identifications of this type will appear with multiple improbable features. The modifications, their locations, and the digestion information should not be considered reliable in these cases, but there will generally be a significant portion of the sequence that is correct, if the confidence is high, which often allows connection to the correct protein or a close homolog. For example, in the annotation in the supplemental data there is a peptide reported as AQCHTVEK with N-terminal carbamylation, deamidation of Gln, carbamidomethylcysteine, and a semitryptic Cys-Ala cleavage at the N terminus. However, a better interpretation of the spectrum would probably be the corresponding tryptic peptide CAQCHTVEK with an internal disulfide, a modification that was not allowed. Most of the sequence is correct, allowing connection to the right protein, but the exact details of the answer are not reliable. Note that this peptide was out of search space for all searches analyzed in the study so this inaccuracy in the annotation had no impact on the results presented here.

Protein Inference Analysis—The peptide ID search results are passed automatically to a third component, the Pro Group Algorithm, which is also part of ProteinPilot Software. This algorithm takes the top 10 peptide hypotheses for each spectrum as input, regardless of maximal confidence, and rigorously distills this set into the set of proteins that can be reported as having been detected with a specified level of confidence in a way consistent with established publication guidelines (33, 34). The Pro Group Algorithm is described elsewhere.[1–3]

User Interface Control and Parameterization—In conventional search engines, all method settings explicitly control how to do the search. In an effort to remove algorithmic complexity and reduce the risk of incorrect parameterization, a user interface was developed that hides virtually all of these direct algorithmic controls. This was achieved by implementing a business logic layer containing a "translation" framework whereby user input can be in the language of the experimental scientist as description of 1) the sample and treatment (cysteine alkylation, digestion, labeling scheme, acquisition instrument, and species) and 2) what is desired from the search in terms of the compromise between speed and the quality of the result. This simple input is then translated into the optimal set of algorithmic settings. For example, selecting trypsin as the digestion agent is translated to a set of digest feature probabilities to capture major and minor specificities of trypsin as well as a background rate for all other potential cleavage sites. This obviates the need to do "semitrypsin" or "no enzyme" searches on a tryptic sample because search space is made large enough for segments with very hot STV that many peptides of these less common types can be identified. The same concept of the translation of workflow factors into more complex feature probability descriptions applies to the rest of the method input. Selecting iodoacetamide as the cysteine alkylation would be translated into a set of feature probabilities that includes the major modification on Cys and also known less frequent side reactions from the reagent. A field called Special factors captures additional workflow steps that would impact how a search should be parameterized. For example, an option in this list called Gel-based ID translates into the increased frequency of oxidation artifact modifications and modifications due to acrylamide. For all sample types, there is a translation that describes the background

rate of general workup (artifact) modifications like pyroglutamic acid formation, oxidations of methionine, and deamidation. The need to set mass tolerances is obviated by inferring expected variances of MS and MS/MS data for the specified instrument. There are only a few actual options about how to do a search. There are ID focus options, which allow the additional consideration of large sets of biological post-translational modifications and/or substitutions. The desired tradeoff between speed and quality of result is indicated by selecting a Rapid or Thorough search. The former only runs the simple Fraglet search component, whereas the latter invokes the Taglet component as well. A sequence database to be searched is selected, and a species constraint function is applied as with other search engines. A table containing all the modification translations is included in the help function within the software, and missing modifications can be defined. A figure showing the method definition screen is included in the supplemental data.

A fully functional trial version of the software is available for download (ProteinPilot, Applied Biosystems). This software is completely independent of the instrument acquisition software so any modern Windows-based computer can run it.

Searches for Comparative Assessment Versus Mascot

A series of five Paragon Algorithm searches and five Mascot searches were run to assess relative performance both between the two search engines and also among the different searches with each engine. Table I shows a summary of the searches run, the parameters used, and measurements from the analysis of each search. Searches were run on two different FASTA format protein databases, the UniProtKB/Swiss-Prot database, which has about 200,000 proteins, and the CDS Combined file (35, 36), which is essentially a version of National Center for Biotechnology Information (NCBI) NR.fas that has been made truly non-redundant for public proteins, and then Celera proteins are added to this set, yielding a total size of about two million proteins. These two were chosen because they differ in size by an order of magnitude, but both carry a full diversity of proteins from different species as was necessary to search the test mixture of proteins from multiple species. Each row in the table represents a type of search where an effort has been made to align similar Mascot and Paragon searches in terms of what peptides can be found by each search. The first row for each database is referred to as the Small Search Space (Small SS) search type. Because the Paragon Rapid search effort setting is like a conventional precursor-type search engine, it is possible to achieve nearly identical search space between the two engines for this search type as indicated by the listed parameters. Custom Paragon modification sets were created, removing the majority of features in the much larger set normally used, to allow exact alignment with Mascot in modifications. All modifications were variable modifications in all of these 10 searches. The only point of difference in the Small SS searches is the mass tolerances as will be discussed later. The second search type will be referred to as Large Search Space (Large SS) search type for which Mascot is run with the semitrypsin digest setting and the Paragon Algorithm is run in its Thorough search effort setting with trypsin specified as the digestion agent. These two searches have been aligned because they both enable finding peptides that only conform to tryptic specificity on one end. There is a substantial amount of these "semitryptic" peptides in the annotation. A few additional

TABLE I

Summary of searches for Paragon-Mascot comparison

Five searches each for Mascot and the Paragon Algorithm are described. Each row contains a pair of searches that are considered the same "search type" for the purpose of comparison. In order of increasing search space, there are three types: Small SS, Large SS, and No Digest SS. Most of these were run on both medium and large databases that differed 10-fold in size. The Correct Full 890 columns give the results of grading the searches against all 890 annotated spectra. Sensitivity measured by total number of right answers in a search in any rank is dominated by the allowed search space for that search, whereas the fraction of those right answers that are ranked as the first answer is a rough measure of discrimination. The First in Shared (Sh.) 397 columns give the results of focusing on only the 397 spectra where all 10 searches were able to find a correct answer in any rank, giving the number of these spectra where the first answer is right and wrong and the correct percentage. The fraction of these spectra where the first answer is correct is a very pure measure of discrimination.

	Paragon Algorithm Search Description	Correct Full 890 First|Total (% first)	First in Sh. 397 Right|Wrong (% correct)	Mascot 2.1 Search Description	Correct Full 890 First|Total (% first)	First in Sh. 397 Right|Wrong (% correct)
UniProt/Swiss-Prot Dec 20, 2005 version 204,086 proteins	**Small SS** *Rapid* mode *Trypsin* (conventional) Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) 2 missed cleavages	448 | 462 97.0%	397 | 0 100%	**Small SS** *Trypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) 2 missed cleavages (Significance = 41)	432 | 456 94.7%	392 | 5 98.7%
	Large SS *Thorough* mode *Trypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl	722 | 742 97.3%	395 | 2 99.5%	**Large SS** *Semitrypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl 3 missed cleavages (Significance = 55)	629 | 680 92.5%	382 | 15 96.2%
	No Digest SS *Thorough* mode *No digest* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl	673 | 724 93.0%	381 | 16 96.0%	**No Digest SS** *No enzyme* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl 3 missed cleavages (Significance = 64)	584 | 671 87.0%	361 | 36 90.9%
CDS Combined Release 5 Mar 2, 2005 1,967,674 proteins	**Small SS** *Rapid* mode *Trypsin* (conventional) Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) 2 missed cleavages	468 | 488 95.9%	394 | 3 99.2%	**Small SS** *Trypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) 2 missed cleavages (Significance = 51)	450 | 488 92.2%	383 | 14 96.5%
	Large SS *Thorough* mode *Trypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl	748 | 784 95.4%	390 | 7 98.2%	**Large SS** *Semitrypsin* Carbamidomethyl (C) Deamidation (N/Q) Oxidation (M) Pyroglu (E) Pyroglu (Q) Protein N-term Acetyl 3 missed cleavages (Significance = 65)	581 | 681 85.3%	342 | 55 86.1%

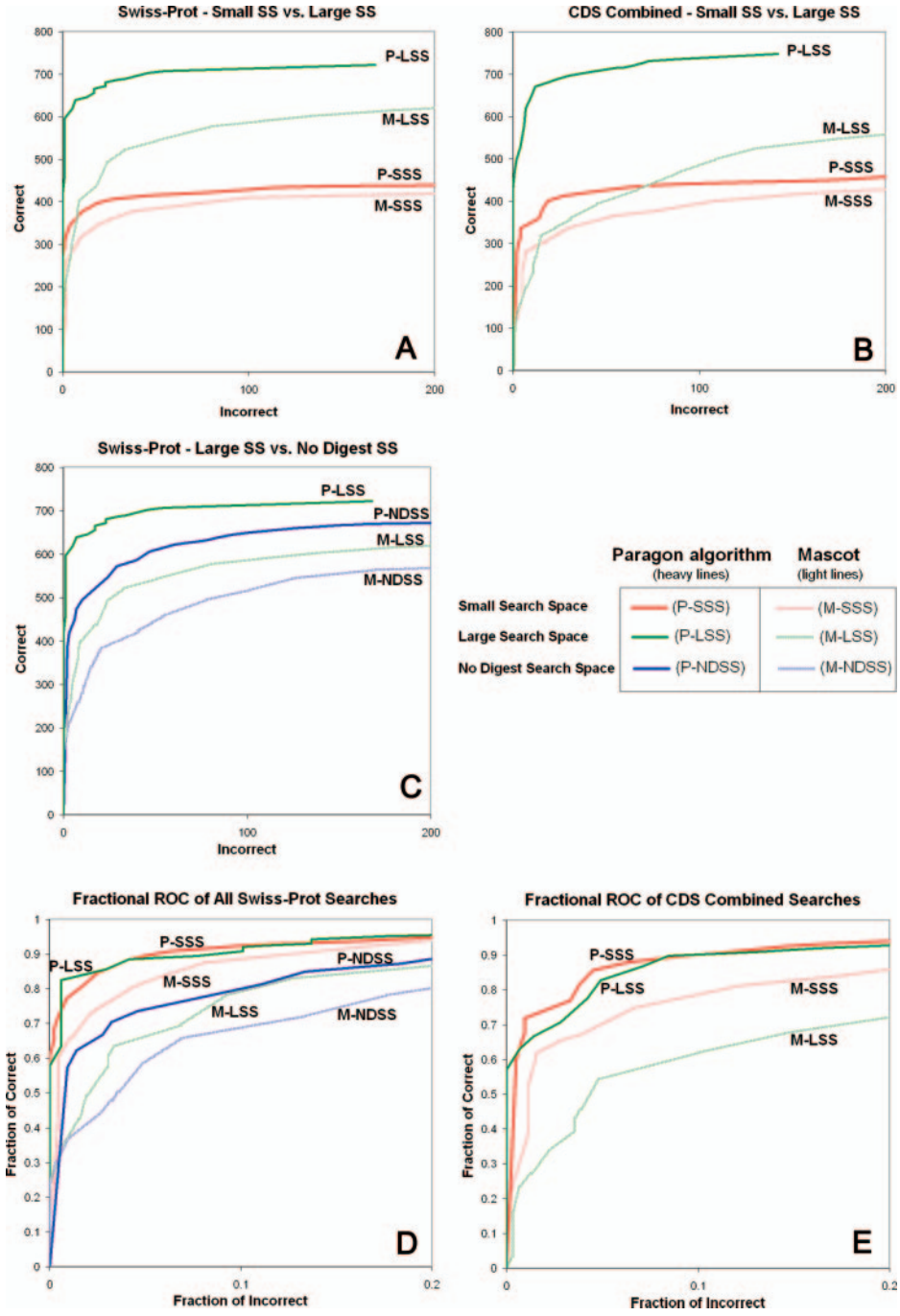


FIG. 2. **ROC curve analyses using all 890 annotated spectra.** The performance of the 10 searches in Table I is evaluated using ROC curves, grading only the 890 spectra that were annotated of the 1987 total in the file. The common legend is presented in the *center*, arranged to parallel the searches listed in Table I. *Heavy lines* are used for all Paragon Algorithm searches, whereas *light lines* are used for all Mascot searches. Color is used to indicate the type of search: *red* for the Small SS search type, *green* for the Large SS search type, and *blue* for the No Digest SS search type. As indicated by the titles, graphs on the *left* are of searches of the UniProtKB/Swiss-Prot database, and graphs on the *right* are searches of the CDS Combined database. *A–C* are numerical ROC plots, measuring the cumulative number of first answers that are correct *versus* the cumulative number of first answers that are graded incorrect as the confidence threshold is decreased. Note that the scale is identical in all three graphs to allow easy comparison. *A* and *B* compare Small SS searches with Large SS searches for the two databases separately. *C* compares the No Digest SS searches with the Large SS searches for UniProtKB/Swiss-Prot. *D* and *E* are conventional fractional ROC plots where the fraction of all cumulative correct first answers of the maximal correct first answers is plotted *versus* the fraction of cumulative incorrect first answers of the maximal incorrect first answers. *D* and *E* present these analyses for the searches of each database separately. All lines in this type of ROC curve start at (0, 0) and end at (1, 1). The *x axis* is enlarged to focus on the critical *top left* region.

highly specific modifications have been added as well. The third search type, referred to as the No Digest Search Space (No Digest SS) search type, has the same parameters as the Large SS type except that all digest conformance requirements were removed, meaning any sequence in the database could be returned. The No Digest SS search was not run on the CDS Combined database because the Mascot search would have taken over a week to run on the hardware used for the comparison, and this size exceeds the RAM per process limitations with 32-bit processing for Paragon, a limitation that will be addressed in future work. The resulting Mascot Significance scores for each search are also listed in Table I.

Analysis of Performance on All 890 Annotated Spectra—The relative performance of the searches in discrimination and sensitivity on the set of 890 annotated spectra was assessed by constructing ROC curves as shown in Fig. 2. The same relative trends were observed separately with both databases as seen by comparing Swiss-Prot results in Fig. 2*A* with the results from searches of the much larger CDS Combined in Fig. 2*B*. The goal of a search engine is to report all right

answers as the first answer with no wrong answers and to have high discrimination allowing it to rank spectra more likely to be correct ahead of spectra less likely to be correct. These curves are one way of measuring how well a search succeeds in doing that. Fig. 2, *A–C*, shows less conventional numerical ROC curves where the ideal result would be a line that runs straight up along the *y* axis to 890, meaning all first reported peptide answers are correct with no errors. One of the most difficult aspects of comparing searches and search engines is separately assessing differences in discrimination (or specificity) *versus* differences in sensitivity. Numerical ROC plots emphasize differences in sensitivity, the absolute number of right answers. For example, it is clear that the larger search space for the Large SS type searches yields many more right answers relative to the Small SS type searches.

The Correct Full 890 columns for each engine in Table I present the number of spectra with correct first answers, the total number of spectra with correct answers in any rank, and the percentage of detected right answers that are ranked first for each of the searches. The total number of correct answers in any rank is mostly controlled by the size of search space and has less to do with discrimination, assuming enough answers are kept per spectrum. The Small SS searches differ between the engines in this measure by only 2.4 and 0%, respectively, for Swiss-Prot and CDS Combined. This means our effort to achieve identical search space came very close. The one boundary on search space that could not be made the same was the mass tolerances. The Mascot Small SS searches were run with 0.15-Da MS and 0.10-Da MS/MS tolerances where the tolerances are applied in mass space. The Paragon Algorithm Small SS uses constant tolerances in *m/z* space, meaning the tolerance in mass space is multiplied by the charge. Given that the totals for detection in the Small SS searches are within a few percent, this difference has little impact. If anything, the small gain in sensitivity for Paragon should come with a relative cost in discrimination. If the search spaces were identical, the *red* ROC curves in Fig. 2, *A* and *B*, would report directly on relative discrimination of the scoring functions between the two search engines. They appear to show a slight advantage for Paragon. To remove some of the effects of this slight difference in search space, we also constructed the more conventional fractional ROC plots, shown in Fig. 2, *D* and *E*, for all searches of the two databases separately. In this type of ROC plot, the differences in the number of spectra with correct first answers are factored out by dividing by the total for each search separately. This is also done for wrong first answers, and thus, all lines start at the origin and end at (1, 1). Comparison of the *red lines* in Fig. 2, *D* and *E*, still suggests there is an advantage in discrimination in the scoring in Paragon. This could be the result of the competitive scoring scheme of Paragon or the use of feature probabilities in scoring. However, the point of running the Small SS type searches on both engines was not at all to show an advantage in scoring but rather to demonstrate that the scoring in the Paragon Algorithm is comparable with that of Mascot and essentially use this search type as control to study the novel algorithmic functions in the Thorough mode searches of Paragon.

The most striking aspects of the results in Fig. 2 are in the comparison of the effects of increasing search space on each search engine. Fig. 2, *A* and *B*, and Table I show that the Mascot Large SS searches yield more correct answers than the Small SS searches; however, this comes with some cost in discrimination. In Fig. 2, *A* and *B*, the *green lines* start to break from the *y axis* sooner than the *red lines* for Mascot. This same tradeoff, increased sensitivity at the cost of decreased discrimination, is not observed with Paragon in going from the Small SS searches to its Large SS searches. Comparison of the *heavy green lines* to the *heavy red lines* in Fig. 2, *A* and *B*, indicates much greater detection with larger search space *without* any apparent loss of discrimination. The *green lines* simultaneously go much higher and clearly break from the *y axis* later than the *heavy red lines*. Although these differences may be difficult to discern in Fig. 2, *A* and *B*, the differences are stark in Fig. 2, *D* and *E*. There is a clear loss of discrimination between the *red* and *green lines* for Mascot, whereas there is almost no detectable difference between the *red* and *green lines* for the Paragon Algorithm despite a huge increase in the effective search space.

Although the sample actually was digested with trypsin, the No Digest SS searches are another important test case for the differences in handling large search space, representing the upper limit in the digestion variable of search space. Table I shows that both Mascot and Paragon lose a few right answers relative to the Large SS searches, 742 down to 724 for Paragon and 680 down to 671 for Mascot. The answers are "lost" because the drastically enlarged search space layers so much statistical noise on top of the signal that the correct answers no longer fall within the top five and 10 reported answers for Paragon and Mascot, respectively. The increase in noise also causes the percentage of right answers in search space that are ranked first to drop from 97.3 to 93.0% for Paragon and from 92.5 to 87.0% for Mascot. Fig. 2*C* indicates the same trend in considering the whole curves rather than just the end points in Table I. Although the Paragon Algorithm does take a hit removing the digest specificity, it is strikingly better than the No Digest SS search of Mascot. Nearly twice as many right first answers are reported by Paragon before the lines begin to break away from the *y* axis. This means the yield of highly confident identifications is approximately double with Paragon for no digest searching. The Paragon No Digest SS search even outperforms the Large SS search of Mascot in Fig. 2*C* and appears to be equal or better in discrimination in Fig. 2*D* (*heavy blue line versus light green line*).

Although we eliminated the modification variable as a source of differences in search space, there are still real differences in total right answers for the two larger search space types listed in the Correct Full 890 columns. To under-

stand what these differences were, we did a detailed examination of the CDS Combined Large SS searches where Paragon found right answers in any rank for 784 spectra compared with 681 for Mascot. A Venn analysis of these searches determined that both search engines found right answers for 677 of the spectra, whereas only Mascot found right answers for four spectra, two where the correct answer was ranked first and two where it was not, and only Paragon found right answers for 107 spectra, 82 of which had correct first answers. The four spectra where only Mascot reported a correct answer all had low confidences (E-values of 130, 12, 4, and 4300) and were all semitryptic peptides. For the Paragon-specific spectra, we focused on the 82 spectra where the first answer was correct. One way search space is larger for Paragon in its Thorough mode is that observed *versus* theoretical peptide delta masses much larger than the tolerances normally used in precursor-type database searches can be considered. This allows good identifications to be recovered when the wrong peak was called as the monoisotopic peak or when a secondary peptide species present within the precursor isolation window contributes to or even dominates the observed fragmentation. To our surprise, the large delta peptides did not account for the majority of additional detections relative to Mascot. 15 of the 82 spectra had delta masses off by close to 1 Da, whereas another 15 had delta masses off by 2 Da or more. About 40% of these 30 correct "large delta" cases had confidences greater than 95%. There were two other differences in search space that could account for some of the remaining 52 spectra in this set of 82. First the number of missed cleavages considered by Paragon is not limited to a fixed value. One peptide had five missed cleavages, accounting for an additional three spectra. Another difference is that the Paragon Thorough mode search with trypsin set as the digestion agent can actually find peptides that do not conform to expected tryptic cleavage on either end, often referred to as "non-tryptic" peptides. Because these are rare, we did not expect this to account for many of the spectra, and accordingly, only six spectra were explained as fully non-tryptic peptides. All of these were verified manually, belonged to the top 11 proteins, and had cohort peptides with overlapping sequence, including semitryptic peptides with common cleavages. In total, differences in search space only accounted for 39 of the 82 spectra, meaning 43 should be in search space for Mascot. Furthermore of the 82 spectra, 74.4% of the correct answers were actually tryptic, 18.3% were semitryptic, and 7.3% were fully non-tryptic. Other than an expected enrichment for non-tryptic peptides, this is essentially the same breakdown as was observed in the whole set of annotated spectra. We manually inspected a sampling of the 43 spectra to see what answers Mascot did report. In many cases, there were so many close alternative sequences that the 10 peptides Mascot saved per spectrum had very little sequence diversity. The right answer was effectively being "pushed below the surface" by the huge amount of wrong answer noise from large search space. To further test this theory, we checked the Small SS Mascot search on Swiss-Prot for these spectra to see whether correct answers could be found and observed that more than half, 24 spectra, did have right answers present, and 16 of these were even ranked first. In other words, the right answers were not being detected for these spectra when searching very large search space because of poor discrimination, not because of differential sensitivity because the allowed search space was different.

Analysis of Performance on the 397 Consensus Spectra—To more rigorously interrogate the relative discrimination performance of the two engines in different search modes, we decided to focus on the subset of spectra where the right answer was within search space for all 10 searches. This means all these spectra had a simple tryptic peptide as the right answer. In this mode of examination, the benefit of larger search space in greater sensitivity (as shown in the study of the full 890 annotated spectra) could only be detrimental to discrimination in this focused examination. As described under "Experimental Procedures," the annotation did contain more answers derived from the Paragon Algorithm than from the Mascot algorithm. By focusing on only spectra where both engines can find the right answer in all modes of search, any negative effects from unintended bias should be removed. For this subset of spectra, the right answer is present for all searches, and thus, comparative analyses report purely on differences in discrimination, directly measuring the impact of increasing "noise" going to larger search space. Of the 890 annotated spectra, 805 had a right answer that was found in at least one of the 10 searches examined, 681 (85% of 805) had right answers in at least six of the 10 searches, and 397 (49% of 805) had right answers in all 10 searches.

We repeated ROC curve analyses using only the 397 consensus spectra. Fig. 3, *A* and *B*, shows the numerical ROC curves for Swiss-Prot and CDS Combined searches, respectively. As would be expected, the discrimination is weaker for any given search on the larger CDS Combined *versus* the same search on Swiss-Prot for all searches with both engines. The First in Shared 397 columns in Table I give the data for the end points of these lines for each engine, listing the number of first answers that are right and wrong and the percentage of the 397 that are right. This percentage is one measure of discrimination.

First let us consider the Small SS type searches (*red lines*). Because the right answer is present in all 397 spectra and the searches are nearly identical in search space between the two algorithms, these curves are reporting directly on differences in discrimination that are due to the scoring function of each engine. The slightly larger search space for Paragon that gave an apparent advantage in Fig. 2, *A* and *B*, can only be detrimental to performance in Fig. 3, *A* and *B*. Nonetheless as was seen in Fig. 2, *D* and *E*, there is still a suggestion that the Paragon scoring discriminates slightly better. Again this search type was intended to be a control, and being very

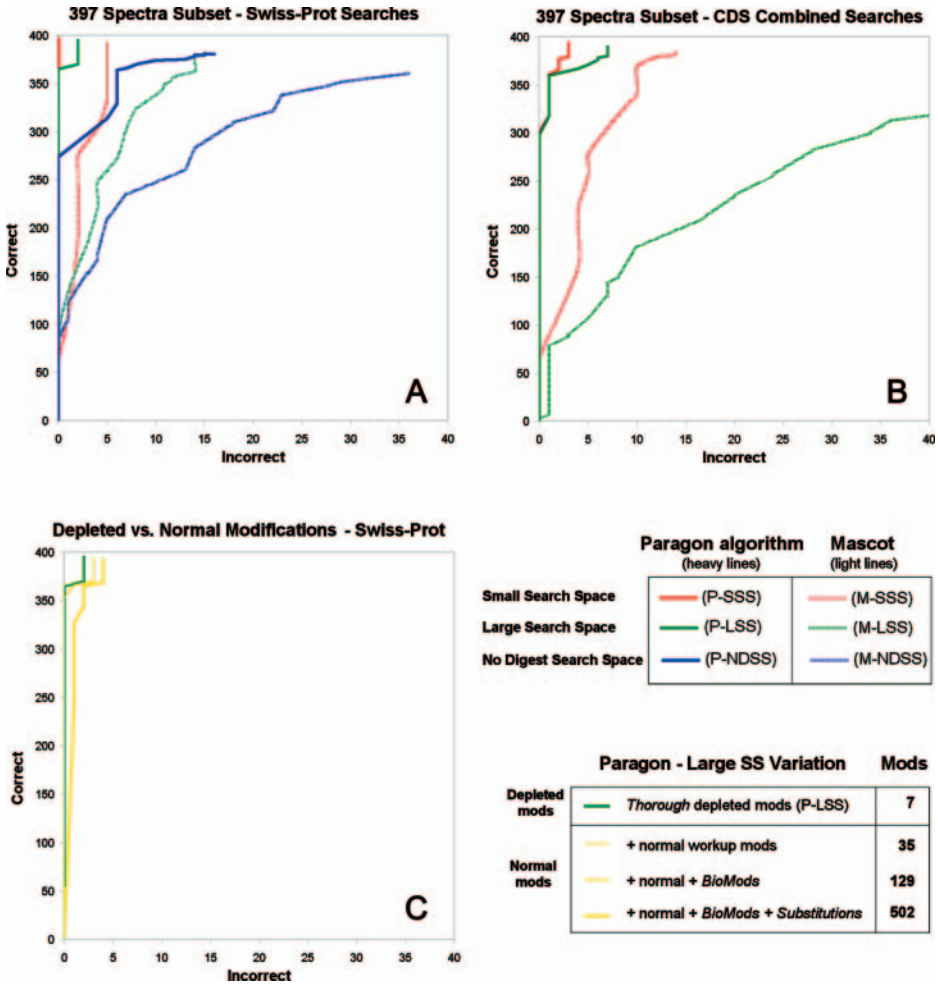


FIG. 3. **ROC curve analyses using 397 consensus spectra.** The formatting to indicate different searches is the same as in Fig. 2, but several additional searches have also been added as listed by the *gold lines* in the *second legend table*. *A* and *B* measure the discrimination of searches on UniProtKB/Swiss-Prot and CDS Combined databases, respectively. Because the right answer can be found in all of the 10 searches in Table I, the right answer for all of these spectra is necessarily a tryptic peptide. Focusing on this subset of spectra measures the effect of increasing noise from moving toward larger search space as it impacts the ability to successfully place the right answer in first place. The differences in the Small SS searches report directly on discrimination differences due to the scoring functions. Considering these differences to be small, the differences in discrimination between the engines in the two large search space types are then entirely due to differences in the number of hypotheses scored. This number is much lower for Paragon Thorough searches because of the use of STV and feature probabilities. *C* shows that there is almost no impact on discrimination with 35, 129, and even 502 modifications, numbers of modification features normally invoked by the settings in the user interface of the Paragon Algorithm. All three graphs are equivalently scaled to ease comparison.

conservative, this difference could be considered to be the margin of error of this study. The main conclusion from the Small SS search comparisons is that the scoring in Paragon is at least on par with the scoring in Mascot in terms of fundamental discrimination.

The most important feature of the ROC results in Fig. 3, *A* and *B*, is the differential impact of increasing search space for each engine. The loss of discrimination going from Small SS to Large SS to No Digest SS (*red* to *green* to *blue lines*) for Mascot is strikingly larger than it is for the same series with the Paragon Algorithm. There is almost no loss of discrimination for Paragon between Small SS and Large SS. As was observed in Fig. 2, *D* and *E*, Fig. 3*A* shows that the Paragon No Digest SS actually discriminates equally if not better compared with Mascot Large SS.

One of the most striking differences between the two engines in analogous cases is the Large SS searches on CDS Combined seen clearly in both the differences between the *green lines* in Fig. 3*B* and the end point data in Table I. Because this was the largest difference between the engines and because it was the largest Mascot search space (having the highest significance threshold), this pair of searches was examined in more detail. In the Mascot Large SS search on CDS Combined, a correct answer that was present in its top 10 hypotheses was not successfully ranked as the first answer for 55 of the 397 spectra (13.9%). For the analogous search with Paragon, the failure rate was only 7 in 397 (1.8%). Both engines failed on five of the same spectra, whereas only Paragon failed on an additional two spectra, and only Mascot failed on an additional 50 spectra. Believing that the main difference in performance between the engines should be because the Paragon Algorithm leverages the additional information from STVs and feature probabilities to score far fewer peptides, we theorized that if we took the reported first answer from the cases where Mascot failed to rank a correct answer first we should find that Paragon did not even score this hypothesis for that same spectrum in many cases. This is exactly what was observed. In 48 of 55 cases, the incorrect answer Mascot ranked as its first answer was not even among the top five hypotheses for the same spectrum for Paragon, meaning it is very likely Paragon did not even score the peptide.

Searches Using Full Modification Sets—Custom depleted modification sets were used for Paragon searches to elimi-

nate the modification variable with respect to Mascot. In normal operation, the Paragon Algorithm actually uses a base-line level of 35 workup modifications in all searches with its Thorough search effort setting and also has user-controllable options to additionally consider a set of 94 biological modifications and/or 376 amino acid substitutions. For clarity, we chose to eliminate the modification variable in the validation of the fundamental new ideas in the Paragon Algorithm. However, as a quick check, Fig. 3*C* shows the same type of analysis on the 397 common set to demonstrate that, even when much larger numbers of modifications are considered, the discrimination still holds up. The figure shows that when considering 35, 129, or even 502 modifications (or substitutions) the discrimination barely decreases. Note, however, that it does change slightly. If the algorithm were using an iterative, filtering approach that removed spectra from further search like many second pass approaches, there would be zero change. However, the separate pass approach in the Paragon Algorithm *is* considering additional hypotheses for these spectra.

Trends in Numbers of Peptide Hypotheses Scored and Search Times—To quantitatively assess differences in the number of peptides that are scored in each search type, we added a counter to the Paragon Algorithm scoring function and exported these data for each spectrum. The median number of hypotheses scored among all spectra was determined for each of the five searches as a measure of the actual search space scored. For Mascot, we determined relative -fold changes in the number of hypotheses scored using the changes in significance threshold among the searches. These results are summarized in Table II. To emphasize the trends more than the absolute numbers, we normalized all search space measures to be described as a relative change over the Swiss-Prot Small SS search for the same engine. Because of this, attention should generally be focused on the relative trends between the searches within the same engine rather than comparing the absolute -fold changes across engines.

The data in Table II show that the -fold increase in hypotheses scored between Swiss-Prot and CDS Combined Small SS searches within each engine was very close to the 9.64-fold that is expected based on the difference in database sizes, 9.05/1.00 and 9.55/1.00 (-fold increase in hypotheses over Swiss-Prot Small SS for CDS Combined Small SS over Swiss-Prot Small SS), for Paragon and Mascot, respectively. However, there is a dramatic difference between the engines in the -fold increase in number of hypotheses scored in going from the Small SS to the Large SS searches. For both databases, the Large SS searches necessitate scoring 25-fold more peptides than the Small SS search for Mascot (25.1/1.00 and 251/10.0), whereas Paragon only scores a very small number of additional peptides between the two search types, about 1.2-fold more in the case of both databases (1.21/1.00 and 10.5/9.05). This supports our previously suggested rationale for the large differences in observed discrimination between the two search engines in the Large SS searches. Notice also that the -fold increase in search space with Swiss-Prot searches from Small SS to No Digest SS for Paragon is very close to the change for Mascot between Small SS and Large SS. Both increases in search space are about 20-fold, which is consistent with our previous observation that the discrimination in the Paragon Large SS and the Mascot No Digest SS searches was comparable. Although these two searches are equal in *scored* search space size as measured by the number of hypotheses scored, they are very different in *effective* search space size. Because the Mascot MS/MS search scores all hypotheses within search space, its scored search space and effective search space are always equal. The use of STVs and feature probabilities allows the Paragon Algorithm to judiciously not score most hypotheses that are allowed in search space, resulting in an actual scored search space size that is much smaller than the effective search space.

Fig. 4*A* estimates the difference between the scored and effective search space size for each of the search types. By plotting the -fold increases for each engine on equivalent searches separately for each search type, we can estimate this difference from the slopes. The *red line* for the Small SS search types yields a slope very close to unity, meaning the effective and scored search space for Paragon is the same because it follows the same trend as Mascot. This is what should be observed because the Paragon Rapid search mode is essentially the same simple precursor-type search that

TABLE II

Number of hypotheses and search times

The trends in the number of hypotheses scored and the time of search are indicated relatively by reporting the -fold increase (ratio) of the number of hypotheses or search time to the corresponding values for the smallest search space search, Swiss-Prot Small SS, separately for each search engine. This emphasizes the trends with increasing search space within each engine rather than the absolute numbers, which are also given in parentheses when they are known. The emphasis on the relative trends is particularly important for search times because the hardware running each search engine was not the same. The significance of the trends in this table are examined in Fig. 4.

	Increase in Hypotheses Scored Fold increase over Swiss-Prot Small SS (median # hypotheses scored)		Increase in Search Time Fold increase over Swiss-Prot Small SS (search time in sec/spectrum*)	
Search	Paragon algorithm	Mascot	Paragon algorithm	Mascot
Swiss-Prot Small SS	1.00 by definition (6739)	1.00 by definition	1.00 by definition (0.124 sec/spect.)	1.00 by definition (0.337 sec/spect.)
Swiss-Prot Large SS	1.21 (8175)	25.1	2.82 (0.351 sec/spect.)	16.6 (5.60 sec/spect.)
Swiss-Prot No Digest SS	19.9 (133,887)	200	40.1 (4.98 sec/spect.)	129 (43.6 sec/spect.)
CDS Combined Small SS	9.05 (61,012)	10.0	7.30 (0.908 sec/spect.)	9.55 (3.22 sec/spect.)
CDS Combined Large SS	10.5 (70,756)	251	12.2 (1.51 sec/spect.)	165 (55.6 sec/spect.)

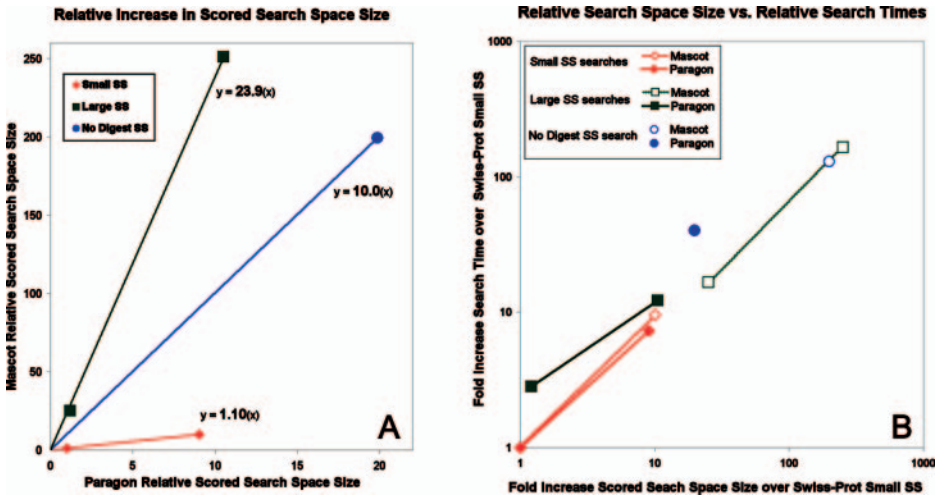


FIG. 4. **Relative trends in number of hypotheses scored and search times.** *A* and *B* visualize the data in Table II to emphasize the trends with increasing search space comparatively between the two search engines. *A* compares the relative increase in the *scored* search space size for each engine as measured by the -fold increase in median number of hypotheses scored over the Swiss-Prot Small SS search for the same engine. The analogous results for the three search types are plotted *versus* each other, and the trend for each type is roughly indicated by fitting a line to $y = m(x)$. The slopes resulting from these fits are shown on the plots. The slopes indicate the relative burden of that search type comparatively between Paragon and Mascot where a slope of less than 1 would indicate the search type costs more work in hypotheses scored for the Paragon Algorithm, whereas a slope of greater than 1 indicates a greater cost for Mascot. The Small SS search type is seen to have a slope of about 1 indicating equal cost, whereas the No Digest SS type costs Mascot ~10-fold more in scored search space size than Paragon, and the Large SS type costs Mascot about 24-fold more than Paragon. *B* connects these effects to the resulting search times. The Small SS searches for the two engines, as indicated by the *red lines*, fall in the same regime of time and scored search space as expected. The Large SS searches, as indicated by the *green lines*, fall in completely different regimes of both scored search space size and search time cost. The single No Digest SS measures are also shifted by an order of magnitude in both axes.

makes no use of sequence tags. The *green line* comparing the Large SS searches shows a dramatic difference with a slope of about 24. This means the Paragon Large SS search has a scored search space that is about 24 times smaller than the effective search space. This assumes that the effective search space of Paragon is the same as the scored search space of Mascot. Because the Paragon search can find fully non-tryptic peptides, very large mass delta peptides, and many more missed cleavages than were allowed in the Mascot search, the Paragon effective search space is actually much larger than this. Thus, this 24-fold estimate is a very conservative lower bound. The *blue line* for the No Digest SS searches has a slightly softer slope around 10 but still much greater than unity, also indicating a large difference between scored and effective search space. The actual scored search space size for the No Digest SS search cannot be reduced as much for the Large SS search because, without any expected digest specificity, all cleavages are treated as equally likely. By contrast, in the Large SS search, telling the Paragon Algorithm that the sample was digested with trypsin invokes a very complete description of the probabilities of cleavages between pairs of residues. To a very rough approximation, the gain from the *red* to *blue lines* is due mostly to the use of tags and STV, whereas the gain from the *blue* to *green line* is due mostly to the use of digest feature probabilities. That is, the difference between Small SS and Large SS for the Paragon Algorithm is due to both the use of STV and digest probabilities, whereas the change from its Large SS to No Digest SS is due to the removal of digest feature probabilities. Because the modifications are constant in the two larger search space search types, the difference between the *blue* and *green lines* reflects the value of having the digest information over not having it.

Up to this point, we have focused entirely on the quality of results in discrimination and detection. Reducing the number of hypotheses scored also yields a large advantage in search time. Table II also presents search times for the 10 searches and scales these relative to the Swiss-Prot Small SS search time within each engine as was done with the number of hypotheses scored. This was particularly necessary for the search times to normalize for differences in the hardware that was running each search engine. Again the point is the relative trends among search types within each engine, not the absolute search times. Fig. 4*B* shows the similarity between the trends in actual scored search space size and the trends in search time. All Mascot searches fall along the diagonal slope of unity through the center of the graph, meaning the increase in search time is directly proportional to the increase in the amount of scoring. Whereas the Paragon Small SS line in *red* falls right on top of the Mascot Small SS line, there is a large difference between the *green lines* for the two engines. The Paragon Large SS searches have effective search space considerably greater than that of those for Mascot, yet the scored search space, as measured by the number of hypotheses

scored, is more than an order of magnitude smaller as shown in 4*A*. This comes with a nearly proportional drop in search time as well as shown in 4*B*. The Paragon Large SS searches take only 2.8- and 1.8-fold more time than the corresponding Small SS searches of the same database for Swiss-Prot and CDS Combined, respectively. The same comparison with Mascot finds a difference of about 17-fold between these two search types for both databases. The reason for the slight vertical shift of the Paragon *green* and *blue points* in Fig. 4*B* is the additional overhead in calling sequence tags, calculating STVs, etc. Based on the improvements in discrimination, sensitivity, and search time that allow the exploration of much larger search spaces, it seems clear this is an overhead worth paying.

DISCUSSION

The central concept put forth with this new search engine is the expansion of search space commensurate with the degree that a segment of the sequence database is implicated by sequence tag evidence. To our knowledge, this is the first example of an algorithm that searches different areas of a database to different degrees on a continuum during the search of a single spectrum. That is, the allowed search space can be different for each sequence segment.

The Paragon Algorithm assesses this degree of implication on a continuous scale that is conceptually referred to as a Sequence Temperature Value. This value is derived by calling many small sequence tags for an MS/MS spectrum with associated estimates of correctness and determining their net effect for each region of the database. The corresponding modulation of search space is accomplished using feature probabilities. Thus, for a segment in the database that is hot for a particular spectrum, *i.e.* strongly implicated by the tag set called for that spectrum, the algorithm will consider peptides with rare modifications, unexpected cleavages, less likely substitutions, and large delta masses. At the other limit for a search of the same spectrum with the same set of tags, a different segment in the database may be very "cold," *i.e.* not at all implicated by the tags, and the algorithm will only consider the mostly likely features or lack of features, for example, only tryptic peptides and the expected cysteine alkylation modification but not its absence or any side reactions.

An alternate way to state the fundamental Paragon concept would be to say that peptide features should be considered such that unlikely peptides are only considered when there is a compensating amount of fragmentation evidence that could substantiate an otherwise improbable answer. One limitation of the Paragon Algorithm is that it may fail to find some peptides that are both low frequency types of peptides (atypical) and have poor fragmentation. This is a deliberate sacrifice to gain the speed and discrimination that has been demonstrated under "Results." The detailed examination of the answers found by only one of the two search engines for the CDS Combined Large SS searches showed a good example of this effect. The four spectra where Mascot found the right answer and the Paragon Algorithm did not all fit this pattern; the right answers were all lower probability semitryptic peptides, and the spectra had poor fragmentation. Peptide identifications of this kind have limited value for protein identification or as peptide results because they are both improbable and lack the spectral information that would be needed to substantiate an improbable answer. They are essentially just peptide mass mapping results.

The results in this study demonstrate that the Thorough search mode of the Paragon Algorithm, which invokes the novel functionality described here, achieves a large increase in search space without the detrimental effects that are typically associated with it. The identification rate is greatly increased, yet discrimination is maintained at almost the same level as small search space. This is possible because there is only a very modest increase in the number of additional peptide hypotheses that are scored relative to small search space, 1.2-fold in the case of searching with tryptic specificity and 20-fold when searching without digest specificity. By contrast, Mascot searches chosen to mimic similar large search spaces showed increases in numbers of hypotheses scored of 25-fold and 200-fold, respectively, for semitryptic and no enzyme specificity searches. Considering the three fundamental concerns for an identification algorithm, sensitivity (search space), discrimination (specificity), and speed, the Paragon Algorithm is not an alternate balancing of these concerns. It yields large gains in all three.

Although the problem of poor discrimination in large search space is theoretically solvable with advanced scoring techniques that introduce feature probabilities during ion scoring, the cost in additional computational time would be large. The Paragon Algorithm STV tag method provides a shortcut to approximately the same solution without paying a high computational price.

To limit the scope of this study, the issue of searching for large numbers of modifications was intentionally avoided. Thus, Paragon custom modification sets were made that would allow exactly equal modifications to be searched, reducing differences in search space with respect to Mascot to only digestion and precursor mass tolerance variables. Normal Thorough searching with the Paragon Algorithm invokes much larger sets of modifications, which can yield identifications of less common and even rare modifications and substitutions. A future publication will explore the identification of atypical peptides and its impact on increasing the fraction of spectra explained. However, it has at least been demonstrated here that these searches have virtually no impact on the discrimination in spectra where typical peptides are the best answer. Searching with the normal Paragon modifications sets considering 35, 129, and 503 modifications and substitutions in Fig. 3*C* showed almost no change in discrimination. However, the fact that the results are not exactly identical proves that the algorithm is not filtering out spectra

that match tryptic peptides as is used in many second pass approaches. The separate pass Taglet search is still considering new answers for these spectra, meaning better answers can still be found, yet there is essentially no cost in discrimination. As with the smaller modification sets, this is because the use of sequence tags to determine STVs and the use of feature probabilities allow the algorithm to be extremely judicious about what additional answers it considers for scoring.

One of the most compelling advances the Paragon Algorithm offers over existing approaches is searching without digestion specificity, an extreme in large search space. Here a somewhat artificial situation was examined where the sample was actually digested with trypsin, but this is a very good test case to measure the performance because it is easy to determine what the right answers should be. The increase in both speed and discrimination over Mascot is large enough that this may open up certain areas of research that have not been tractable for lack of a good analysis method. This may include the search for biomarkers in endogenous peptide or "peptidome" samples (6–9), study of neuropeptides (5), and immunology research (10).

It is important to note that, although the No Digest SS search type has been included in this study as a test for validation, you would never run a Paragon Algorithm no digest search unless there was really no regular digestion in the sample. This is counter to the use of conventional search approaches where iterative or filtering approaches often methodically relax digest constraints to identify more peptides in a sample that actually was treated with a digestion agent like trypsin. The Thorough search effort of Paragon with trypsin indicated as the digestion agent finds all cleavage variants directly without the typical costs in loss of discrimination and without the complexity of creating a multistep search.

Paragon offers a number of advantages over a class of methods commonly referred to as "second pass searching," which have become a popular solution for increasing the fraction of spectra identified. There are many different variants of this approach, and they are not necessarily limited to two passes. What all of them have in common is an initial search followed by the application of a filter to remove the great majority of proteins from consideration in subsequent round(s) of the search where search space is then increased by various means such as increasing the number of modifications considered and missed cleavages allowed and relaxing the requirement of conformance to the expected digestion pattern. The Mascot error-tolerant search (37) allows the selection of only a handful of proteins that are assumed to be correct after the first pass, whereas Phenyx (23) can be applied to a full list of proteins identified in a first round. X!Tandem (21, 38) can perform multiple rounds of refinement, and tools like the Mascot Daemon (Matrix Science) allow the user to define complex iterative strategies to do things like filter spectra that have been identified with sufficiently high confidence and search multiple databases.

There are several problems with these approaches, the largest of which is the sensitivity to the protein threshold used after the first stage. If the threshold is too high, then valid proteins (and additional peptides from those proteins) are lost. If the threshold is too low, then false proteins are included in subsequent passes. Searching false proteins and allowing for many modifications and other atypical features can only introduce false peptide identifications. A more subtle risk of applying protein thresholds arises from protein inference complexities stemming from the presence of equivalent or nearly equivalent protein entries in databases. Some of this redundancy is purely informatic, arising from redundant entries and errors in entries, but there is also true biological complexity in protein homologs, splice variants, mutations, etc. Overly aggressive filtering of nearly redundant proteins in the first stage can preclude the detection of additional variant forms or refinement of which form is being detected. By contrast, Paragon does not set a protein threshold; rather it uses continuous probabilities to describe what is learned from an initial search, thereby circumventing all of the problems discussed above.

The second major problem with second pass approaches is that they are not applicable as a strategy for efficient search to cases where it is not possible to do a fast initial search. This is true of samples of endogenous peptides, which lack a regular digest pattern, for example. The first pass search of this type of sample must be done in no enzyme or No Digest SS mode, and thus, the first pass with conventional software is neither fast nor highly discriminating. Paragon addresses this problem by performing a tag-based search (Taglet component) rather than a precursor-filtered search as the initial search.

Finally second pass approaches involve multiple steps and thus are very user-guided and inherently harder to use regardless of the user's level of expertise. The average biologist who is not an expert in informatics cannot develop valid complex methods. Minimally the results across multiple users will be highly variable and hard to validate or judge. For experts, the flexibility and large number of parameters for these tools allow rapid prototyping of different search strategies, but for non-experts this becomes a burden, and the heuristic rules they invent are likely to offset the elegance in the fundamental scoring algorithm of the tool. It must be recognized that virtually all of the tools used in this field today are sufficiently difficult for non-experts in informatics or mass spectrometry and that this is one of the main factors inhibiting the growth of mass spectrometry-based proteomics. The dual use of feature probabilities for both algorithmic purposes and the simplification of the user interface achieves extensive identification like multipass approaches but with a great reduction in the complexity of operation.

Although this work has used only QqTOF data, our experiences so far indicate that the benefits of the algorithm are quite general to other types of tandem mass spectral data,

and accordingly the software also supports the analysis of data from other instruments such as TOF-TOF and ion traps.

In conclusion, the Paragon Algorithm is shown to represent a substantial advance for protein identification by mass spectrometry. The performance advances enable searching large search space as common practice and may popularize some less traveled work flows such as the study of endogenous peptides. Although the advances in the ease of use may be less interesting to mass spectrometry gurus, it should be recognized that software must be easier to use for proteomics to transition from the realm of gurus to the laboratories of biologists. The Paragon Algorithm can be a solid step in this direction.

Acknowledgments—We thank Marjorie Minkoff of Applied Biosystems for the preparation of the protein mixture examined in this study. We thank Lauren Mansfield, Winnie Leung, Liliya Shilova, and Vera Loboda for software testing; Jim Bohannon, Bret Pehrson, Robert Deutschman, Brennan McBride, Cathy Frantz, and Natalia Belyaeva for additional engineering support; and Lisa Schaechter for technical writing. This study would not have been possible without the hard work of this whole team.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

‡ Both authors made equal contributions to this work.

§ To whom correspondence should be addressed: Applied Biosystems/MDS Sciex, 850 Lincoln Centre Dr., Foster City, CA 94404. Tel.: 510-708-9483; Fax: 650-638-6223; E-mail: seymousl@appliedbiosystems.com.

¶ Present address: Molecular Devices Corp., Union City, CA 94587.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Sadygov, R. G., Coriorva, D., and Yates, J. R. (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1,** 195–202
3. Kapp, E. A., Schütz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., and Simpson, R. J. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* **5,** 3475–3490
4. Nesvizhskii, A., and Aebersold, R. (2005) Interpretation of shotgun proteomic data. *Mol. Cell. Proteomics* **4,** 1419–1440
5. Fricker, L. D., Lim, J., Pan, H., Che, F.-Y. (2006) Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrom. Rev.* **25,** 327–344
6. Hardt, M., Thomas, L. R., Dixon, S. E., Newport, G., Agabian, N., Prakobphol, A., Hall, S. C., Witkowska, H. E., and Fisher, S. J. (2005) Toward defining the human parotid gland salivary proteome and peptidome: identification and characterization using 2D SDS-PAGE, ultrafiltration, HPLC, and mass spectrometry. *Biochemistry* **44,** 2885–2899
7. Hardt, M., Witkowska, H. E., Webb, S., Thomas, L. R., Dixon, S. E., Hall, S. C., and Fisher, S. J. (2005) Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal. Chem.* **77,** 4947–4954
8. Geho, D. H., Liotta, L. A., Petricoin, E. F., Zhao, W., and Araujo, R. P. (2006) The amplified peptidome: the new treasure chest of candidate biomarkers. *Curr. Opin. Chem. Biol.* **10,** 50–55
9. Villanueva, J., Shaffer, D. R., Philip, J., Chaparro, C. A., Erdjument-Bromage, H., Olshen, A. B., Fleisher, M., Lilja, H., Brogi, E., Boyd, J., Sanchez-Carbayo, M., Holland, E. C., Cordon-Cardo, C., Scher H. I., and Tempst, P. (2006) Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Investig.* **116,** 271–284
10. Purcell, A. W., and Gorman, J. J. (2004) Immunoproteomics: mass spectrometry-based methods to study the targets of the immune response. *Mol. Cell. Proteomics* **3,** 193–208
11. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
12. Pappin, D. J. C. (1994) Chemistry, mass spectrometry and peptide-mass databases: evolution of methods for the rapid identification and mapping of cellular proteins, in *3rd International Symposium on Mass Spectrometry in the Health & Life Sciences*, (Burlingame, A. L., and Carr, S. A., eds) *San Francisco, September 13–18, 1994*, Humana Press, Clifton, NJ
13. Tabb, D. L., Saraf, A., and Yates, J. R. (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **75,** 6415–6421
14. Tanner, S., Shu, H., Frank, A., Wang, L., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639
15. Taylor, J. A., and Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11,** 1067–1075
16. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567
17. Clauser, K. R., Baker, P. R., and Burlingame, A. L. (1999) Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882
18. Eng, J., McCormack, A. L., and Yates, J. R., III (1994) 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
19. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
20. Bafna, V., and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **17,** Suppl. 1, S13–S21
21. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467
22. Field, H. I., Fenyö, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47
23. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–1463
24. Tang, W. H., Halpern, B. R., Shilov, I. V., Seymour, S. L., Keating, S. P., Loboda, A., Patel, A. A., Schaeffer, D. A., and Nuwaysir, L. M. (2005) Discovering known and unanticipated protein modifications using MS/MS database searching. *Anal. Chem.* **77,** 3931–3946
25. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L., (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting quadrupole collision cell, time-of-flight mass spectrometer. II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell. Proteomics* **4,** 1194–1204
26. Pappin, D. J. C., Rahman, D., Hansen, H. F., Bartlet-Jones, M., Jeffery, W., and Bleasby, A. J. (1996) Chemistry, mass spectrometry and peptide-mass databases: evolution of methods for the rapid identification and mapping of cellular proteins, in *Mass Spectrometry in the Biological Sciences*, pp. 135–150, Humana Press, Totowa, NJ
27. Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight Mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926
28. Liska, A. J., Sunyaev, S., Shilov, I. V, Schaeffer, D. A., and Shevchenko, A.

(2005) Error-tolerant EST database searches by tandem mass spectrometry and multiTag software. *Proteomics* **5,** 4118–4122
29. Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., and Shevchenko, A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* **75,** 1307–1315
30. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, D., and Lajoie, G. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17,** 2337–2342
31. von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X.-j., Goodlet, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to the quantitative protein profiling via ICAT and tandem mass spectrometry: I. Statistically annotated data sets for peptide sequences and proteins identified via the application of ICAT and tandem mass spectrometry to proteins copurifying with T cell lipid rafts. *Mol. Cell. Proteomics* **2,** 426–427
32. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzihrahszky, K. F., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting quadrupole collision cell, time-of-flight mass spectrometer. I. How much of the data is theoretically interpretable by search engines? *Mol. Cell. Proteomics* **4,** 1189–1193
33. Bradshaw, R. A., Burlingame, A. L., Carr, S., and Aebersold, R. (2006) Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* **5,** 787–788
34. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3,** 531–533
35. Kerlavage, A., Bonazzi, V., di Tommaso, M., Lawrence, C., Li, P., Mayberry, F., Mural, R., Nodell, M., Yandell, M., Zhang, J., and Thomas, P. (2002) The Celera Discovery System. *Nucleic Acids Res.* **30,** 129–136
36. Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., Vandergriff, J. A., and Doremieux, O. (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31,** 334–341
37. Creasy, D. M., and Cottrell, J. S. (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2,** 1426–1434
38. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316