Meeting Summary

NCI/NIST Workshop for Next-Generation Methods of Peptide Identification

November 15, 2007

National Institute of Standards and Technology Campus
Building 227, Room 202
Gaithersburg, Maryland

National Cancer Institute
National Institutes of Health
U.S. Department of Health and Human Services

National Institute of Standards and Technology

Contents

Welcome and Overview **1**
Henry Rodriguez, Ph.D., M.B.A.; Stephen Stein, Ph.D.; Christopher Kinsinger, Ph.D.

Plenary Lecture: Proteomics 2.0—Better, Faster, Cheaper **2**
Ronald Beavis, Ph.D.

Identification Methods Session **3**

Refining Peptide Fragmentation Models for Improved Confidence in Sequence/ Spectrum Matching **3**
Karl Clauser, Ph.D.

Roles for Sequence Tagging in Protein Identification **4**
David Tabb, Ph.D.

Physical Models for Predicting Peptide Fragmentation Spectra **6**
Zhongqi Zhang, Ph.D.

Spectrum Library Building and Searching **7**
Henry Lam, Ph.D.

Assigning Confidence to a Peptide ID **8**
Alexey Nesvizhskii, Ph.D.

Combining MS/MS Sequence and Spectral Library Search Algorithms **9**
Lewis Geer, Ph.D.

User Community Session **10**

Perspective and Progress in Protein and Peptide Identification **10**
Sean Seymour, Ph.D.

Getting Your Software Out There **11**
Brian Searle, Ph.D.

Applications for Open-Source Software in Proteomics **11**
Philip Andrews, Ph.D.

Next Steps and Closing Remarks **12**

Welcome and Overview

Henry Rodriguez, Ph.D., M.B.A., National Cancer Institute; Stephen Stein, Ph.D., National Institute of Standards and Technology; Christopher Kinsinger, Ph.D., National Cancer Institute

Dr. Kinsinger welcomed attendees and thanked them for their participation and enthusiasm regarding improving the reliability of peptide identification. Dr. Stein also welcomed attendees, commenting that the scientific community is becoming more detailed and molecular in its thinking. He stated that the National Institute of Standards and Technology (NIST) has worked to provide standards for measurement for more than 100 years and noted that measurements must be reproducible within error limits to foster the communication that is the cornerstone of the research enterprise.

On behalf of the Director and Deputy Director of the National Cancer Institute (NCI), Dr. Rodriguez welcomed attendees to this NCI/NIST jointly sponsored workshop. He noted that the Human Genome Project, which was completed in 2001, represented a touchstone example of a successful, large-scale joint partnership between Federal agencies and the private sector. In this effort, standards were implemented to support a body of research. He observed, however, that the utility of the genomic data for medicine is still pending. DNA is more of a recipe than a blueprint—to capitalize fully on its potential, researchers must understand the underlying biology and interactions of metabolites. In comparison, proteomics represents an exponential increase in technological complexity. The number of functional human peptides/proteins in the human proteome has been estimated to be between 10^5 and 10^6, plus an unknown number of modified proteins (e.g., those altered by alternative splicing, enzymatic alterations, and posttranslational modifications [PTMs]). Moreover, Dr. Rodriguez observed that proteomics has no equivalent to polymerase-chain reaction (PCR) technology; multiple and integrated technologies will be required to capture the proteome's complexity. However, protein database submissions, although still fewer than genomics, are beginning to increase exponentially, leading researchers to ask which are relevant to disease. Although Polanski and Anderson (*Nat Rev Cancer* 2005;5:845-856) identified 1,261 putative protein/peptide biomarkers described in the literature, only 9 of these are currently approved as "tumor-associated antigens" by the U.S. Food and Drug Administration (FDA).

Why are there so few validated protein biomarkers? Discovery output suggests that cancer biomarkers exist, although there are difficulties translating them to the clinic. Given the complexity of the proteome and limitations of current technology, Dr. Rodriguez noted that there are undoubtedly many proteins and peptides yet undetected. Although a high number of biomarker "discoveries" have been reported, a lack of validation indicates a pipeline issue between the phases of discovery and clinical translation.

Many sources of variability exist in proteomics, including experimental study/design, platform evaluation, specimen-related procedures, data acquisition/bioinformatics, data analysis, and data reporting. If proteomics technologies are to be translated into clinical diagnostics, universally accepted metrics will be needed at many steps along the way to help clarify and make comparable experimental results and protocols. Dr. Rodriguez noted that the NCI has initiated the Clinical Proteomic Technologies for Cancer (CPTC) initiative to drive technology optimization and development. The CPTC is an integrated approach to addressing barriers in the biomarker pipeline. The initiative will develop metrics to address variables and provide the scientific community with a foundation of technologies, reagents and standards, data, and analysis systems.

Dr. Rodriguez concluded his introduction by stating that he looked forward to a lively and candid discussion. He noted that suggestions made at this workshop will be presented to the NCI.

Plenary Lecture: Proteomics 2.0—Better, Faster, Cheaper

Ronald Beavis, Ph.D., University of British Columbia

Dr. Beavis began by noting that proteomics is now being seen as a desktop enterprise; the field is undergoing a transition from discovery/development-based technology to clinical, static instruments and methods. As it is currently practiced, proteomics works well, especially when compared to a decade earlier, when it was not certain that rigorous protein identifications could be made. He noted also that any mainstream, conventional search engine will yield similar identifications for abundant proteins in the hands of a sophisticated user. Moreover, the instrumentation has consistently improved, providing better and more reproducible data for informatics analysis. Protein sequence collections have also become increasingly accurate and better integrated with genomic and transcriptomic information. Additionally, the number and range of genomes that have been characterized have increased. Finally, the ability of experimental groups to conduct and perform proteomics experiments has improved steadily; biologists now have a more sophisticated understanding of how proteomic technologies can be deployed.

Dr. Beavis noted that the core elements of the basic experiment to identify proteins (e.g., digestion, fragmentation, analysis) have remained relatively constant. Nonetheless, the public remains largely unaware of how search engines work. However, the process is beginning to converge; a large percentage of genes with ontology codes have been associated with cellular locale, and a large percentage of proteins have been identified that are associated with cellular structural elements. The proteomic data user community is dispersed throughout the country and the world. Also, most proteins show very reproducible peptide patterns. Multiparameter models to predict fragmentation patterns observed from mass spectrometric (MS) analysis are illuminating the physical chemistry of ionization and fragmentation. Also, annotated spectrum libraries (the percentage of residues covered is approximately 21 percent) can be created that are informed by homology of closely related species. Furthermore, spectra can now be arrayed in computer memory and compared with observed data for purposes of matching. Proteomics information can now be mapped to the transcriptome and genome.

However, several aspects of proteomics have yet to be optimized, including laboratory automation, standardization of informatics, accession numbers, and the equation of analysis quality with the number of peptides identified. With regard to laboratory automation, it was anticipated that proteomics would follow the same model as genomic sequencing, in which a small number of highly automated laboratories would produce all of the necessary proteomic data. However, writing and customizing the appropriate laboratory information management systems has proved difficult. Moreover, based on the original aspirations of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO-PSI), there should be standard extensible markup language (XML) formats that allow the interchange of raw and processed experimental data and conditions. Moreover, the informatics analysis should also be in a consistent format. However, these issues have proved challenging due to variability in instrumentation and reluctance from standards committees to generate file formats that effectively capture domain-specific information. In addition, protein accession numbers should provide a stable method for reporting protein identification experimental results; however, these numbers change with time—a fraction of them simply vanishes each year. For example, an examination of a 2003 paper on platelet proteomics shows that one-third of the reported accession numbers correspond to protein sequences that are no longer available. Finally, in an effort to demonstrate the effectiveness of proteomics techniques, the quality of analysis is often measured by the number of identifications rather than the statistical significance of the results. Dr. Beavis observed that this development may reflect a perceived competition with cDNA chip technologies or initial overoptimism.

Dr. Beavis also noted that several developments will be needed over the next 5 years. First, the U.S. scientific community will require data- and results-oriented resources. Current resources include the European Bioinformatics Institute's (EBI) Proteomics Identifications (PRIDE) database (not supported by

the United States), Institute for Systems Biology (ISB) PeptideAtlas (not supported by ongoing funding), and Global Proteome Machine Database (GPMDB). All three databases have met their stated goals, and it is reasonable to ask funding agencies to develop programs to provide long-term support for these types of data archives. However, a coordinated funding program must also reflect the lifecycle of computational efforts. Viable options include P41 grants, which are large, stable center grants that can serve as resources for algorithm development, implementation, and data storage; R21 awards, which allow demonstration of the feasibility of new ideas and algorithms in proteomics bioinformatics; and Small Business Innovation Research/Small Business Technology Transfer (SBIR/STTR) awards, which support small businesses that are developing new applications for commercial use.

Dr. Beavis concluded by noting several additional issues that will affect future proteomics efforts. For example, successful integration of annotated spectrum library analyses into laboratory informatics strategies will require improved methods for global and local library curation and informatics platforms that can handle real-time applications. Also, user interfaces that require modern Web technologies such as dynamic hypertext markup language (HTML), asynchronous JavaScript and XML (AJAX), and wikis and folksonomies must be improved. In general, the routine integration of prior knowledge into displays of results must be improved. Also, improved integration between unique boutique proteomics databases is necessary; this interaction requires protein-specific accession number disambiguation and stable, published query interfaces for public data resources. Improvement of data handling and manipulation of large datasets will similarly require commonly available, aggressive lossy compression strategies (e.g., an mp3 file format for proteomics) and common acceptance and demonstrated long-term viability of data repositories (e.g., Tranche, PeptideAtlas).

Identification Methods Session

Refining Peptide Fragmentation Models for Improved Confidence in Sequence/Spectrum Matching

Karl Clauser, Ph.D., Broad Institute of MIT and Harvard

Dr. Clauser began by observing that tandem mass spectrometric (MS/MS) spectral interpretation strategies exist along a continuum that spans from de novo methods to the searching of spectral libraries. Whereas the former strategy can identify unexpected PTMs and multiple mutations, the latter approach is appropriate for replicate sample analysis and to identify related spectra. He speculated that either spectral library or sequence database approaches will be used in future experiments. Success of these approaches depends on numerous factors, including the enzyme used for protein digestion, the instrument model, and the collision energy. However, the naive fragment ion models in current database search programs tend to substantially overpredict the number of fragment ions (i.e., allow for all possible ion types with similar intensities). By contrast, SpectrumMill's core scoring of MS/MS interpretations uses an assignment bonus (weighted by ion type) minus the nonassignment penalty (weighted by intensity). The goal of this approach is to increase the gap between the scores of correct and incorrect sequences for each spectrum. An additional bonus is given to those dominant peaks that should be dominant, and dominant peaks are penalized if they are assigned to sites that should not be dominant. Under this rubric, correct peak assignments should rarely be penalized. At a dominant site, formation of *b*- or *y*-type ions depends primarily on the location of basic residues in the peptides, rather than on the fragmentation site.

Dr. Clauser concluded by stating next steps for SpectrumMill, which include:

- Refining definitions of dominant adjacencies and positions by adding or combining factors to increase the frequency of occurrence/expectation
- Enhancing the bonus class-cleavage matrix by refining the intensity ratio and dominant-ion categories across charge states and verifying suitability across multiple enzymes and collision energies
- Testing the scoring system on phosphopeptides and other labeling chemistries that produce strong marker ions and/or bear charge
- Examining diminishing the relative score of minor ion types (e.g., *a*, *b/y*-H_2O, *b/y*-NH_3)
- Minimizing the role of secondary scoring in validation
- Evaluating/integrating the Zhang fragment model (see "Physical Models for Predicting Peptide Fragmentation Spectra" discussion, page 6, for details)

Discussion

One attendee asked whether the SpectrumMill algorithm had been extended into rank-based prediction. Dr. Clauser noted that intensity may be accounted for using a variety of approaches; the SpectrumMill approach focuses on gas-phase chemistry to determine the types of peaks generated. When constant neutral losses (e.g., such as that of water or ammonia) are included, the number of identifications increases tremendously. Another participant inquired about the probability of additional false positives that arise from dominant peptides with dominant peaks in precisely the right places; e.g., if these are given too much weight, can the algorithm actually increase the number of false-positive scores? Dr. Clauser noted that at the moment, the opposite is true; peaks are not weighted sufficiently. He noted also that low-intensity events only minimally boost the score.

Another discussant asked how an extra layer of scoring will affect search algorithms. Dr. Clauser observed that this extra layer of scoring was necessary only for the top level of scoring peptides. The goal is to discard the majority of peptides that can be identified; this approach should not impact searching strategies. A participant asked Dr. Clauser to offer his perspective regarding how these models adapt to electron-transfer dissociation (ETD) or matrix-assisted laser desorption ionization (MALDI) techniques. He expressed some concern but stated that in general, many of the features will be observed on non-ion-trap instruments. While the principles are expected to hold true, the magnitudes will likely differ.

Roles for Sequence Tagging in Protein Identification

David Tabb, Ph.D., Vanderbilt University School of Medicine

Dr. Tabb discussed sequence tagging, which enables additional extraction from datasets. He noted that standard database searching can only identify peptides that can be found in a protein sequence database. Software codes must generate a list of likely peptides, and the number of searches is then reduced by a peptide mass filter. Each spectrum is then compared against 50,000-60,000 sequences for tryptic searches in the human IPI database. However, as more comparisons are made, the probability that the search engine will become saturated increases, and sensitivity will decrease. The match evaluator in each database search algorithm then determines whether predicted fragment ions match those seen on the MS/MS spectrum. In this model, fragment ions produced by the spectrometer are used only once in the database search process. In contrast, sequence tagging uses MS/MS information twice—to infer a sequence and to evaluate a spectral match. Tags are partial sequences; the tag's residues are represented by inter-peak *m/z* differences, and flanking masses are included. However, when a spectrum is dominated by a modification to a precursor ion (e.g., loss of a water molecule), identifying tags within the spectrum can be challenging. Sequences can then be evaluated in the context of a sequence database to determine the remainder of the sequence.

Sequences can be inferred from MS/MS spectra using a spectrum graph-based inference approach (Bartels C et al. *Biomed Environ Mass Spectrom* 1990;19:363-368). This approach applies machine-learning strategies to recognize significant peaks in spectra, and it combines information among related fragment ions to evaluate the probability of breakage at a given mass within the peptide. This information is then used to construct a graph that summarizes the likely peptide breakpoints and to read possible sequences from it. Many tools use spectrum graphs, including Lutefisk (Taylor JA et al. *Rapid Commun Mass Spectrom* 1997;11:1067-1075), PepNovo (Frank A et al. *Anal Chem* 2005;77:964-973), InsPecT (Tanner S et al. *Anal Chem* 2005;77:4626-4639), and NovoHMM (Fischer B et al. *Anal Chem* 2005;77:7265-7273). Direct scoring inference tools, which score sequences directly against observed peaks rather than against a graph abstraction, can also be used. Examples include Peaks (Ma B. *Rapid Commun Mass Spectrom* 2003;17:2337-2342) and GutenTag (Tabb D et al. *Anal Chem* 2003;75:6415-6421). Tags are useful for filtering spectra by quality, enabling full-length sequence inference, and identifying posttranslational modifications and mutations.

Quality filtering (employed by SpectrumMill) selectively removes those spectra that lack reliable tags and reduces spectral collections prior to database searching, thereby speeding the processing. Quality filtering also enables subsequent examination of unidentified spectra that have high tag scores. Dr. Tabb then described "database reconciliation," a process by which tags and databases can be reconciled. First, database peptides are required to contain at least one inferred tag sequence for a spectrum. If both terminal masses match, the peptide can be matched as is. Conversely, if neither terminus matches, then the process halts (this could result in doubly modified peptides failing identification). If only the N-terminus mass matches, the C-terminal mass can be modified to force a fit and vice versa. This strategy identifies peptides bearing single PTMs or mutations.

Dr. Tabb noted that sequence tagging will not replace database searching; rather, information from the two approaches should be combined. Optimizing the way in which these kinds of analyses can be integrated will be an early challenge for sequence tagging. It may be necessary to modify false-discovery rate assessment methodologies to support tag-based interpretation. Of course, improvements in sequence inference accuracy will aid the development of this strategy. Even with successes for all of these advances, the proteomics community will take some convincing before it accepts that sequence tagging is worthwhile for routine application. Demonstrating that this strategy is robust and improving its ease of use may enable the proteomic public to realize the gains possible through sequence tagging.

In summary, Dr. Tabb concluded that sequence tagging may greatly expand the scope of peptides that can be identified. While tagging will not replace database searching, it can complement it. If a fraction of the effort expended on database search scoring were instead devoted to tagging, more would be learned about the data.

Discussion

One discussant asked about multiple mutations in a sequence. Dr. Tabb observed that a sequence inferred from a spectrum may contain mutants, and the MultiTag algorithm has attempted to address this scenario (Sunyaev S et al. *Anal Chem* 2003;75:1307-1315). If modification occurs on both sides of the tag, then the identification process halts as a result of the reduction in sensitivity caused by the excess number of comparisons. Another attendee inquired whether a good database search result and a bad tagging result could both be obtained. Dr. Tabb replied that 80-90 percent of confident spectra generate at least one reliable 3-amino acid-tag (a "three-mer"). He has indexed sequence databases to locate every occurrence of every three-mer tag, but he did not realize a substantial improvement in search speed through this strategy. Dr. Tabb noted that it can be difficult to make a judgment call on whether to keep a spectrum on the basis of a single tag.

Physical Models for Predicting Peptide Fragmentation Spectra

Zhongqi Zhang, Ph.D., Amgen Inc.

Dr. Zhang began by noting that although MS/MS spectra can often provide abundant sequence information, all *b*- and *y*-fragments are not equal. Because protein characterization requires identification of all peptides in a sample, this application calls for informatics capabilities different from those for protein identification. He then discussed the mobile proton model for peptide fragmentation in the gas phase using collision-induced dissociation (CID), in which a proton on the basic residue is excited and moves to the peptide backbone, ultimately producing *b*- and *y*-ions (Wysocki et al. *J Mass Spectrom* 2000;35:1399-1406). This process is thus referred to as "charge-directed" fragmentation. An exception can occur when the backbone cleavage site is near an acid residue. In this instance, the acidic side chain may facilitate the cleavage of the peptide bond. Because a charge is not involved in the fragmentation process, this pathway is called charge-remote fragmentation.

Dr. Zhang observed that peptide fragmentation is a competition between pathways, the result of which can be modeled using classical chemical kinetics (Zhang et al. *Anal Chem* 2004;76:3908-3922; Zhang et al. *Anal Chem* 2005;77:6364-6373). Charge distribution can be calculated using a Boltzmann distribution. An Arrhenius equation can be used to calculate the rate constant for each pathway, and first-order kinetics can be used to calculate fragment abundance. Fragmentation channels considered in this model include charge losses, neutral losses, and backbone cleavages (which can be charge-remote or charge-directed). Major assumptions of the model include:

- The effective temperature is related to the mass, charge, and collision energy.
- The side chain of each residue has a constant GB.
- The GB of a backbone amide depends on the neighboring residues, and the effects are additive.
- Intra-molecular proton transfer proceeds rapidly.
- Activation energy for each backbone cleavage depends on the neighboring residues, and these effects are additive.
- The same fragmentation mechanism has the same activation factor.

This model predicts ions created by neutral loss, backbone cleavage, and C-terminal rearrangement. Ions predicted within two generations include a range of immonium ions, internal ions, *a*-ions, truncated peptides, and additional *b*- and *y*-ions, among others. The current version of the model contains 336 parameters (15 of which are instrument-dependent). Initial values were from best estimates based on literature values. This model was optimized using more than 10,000 high-quality ion-trap spectra. A similar physical model is used to predict peptide ETD spectra.

The model can process about 100 spectra per second. Issues with the model include computation speed (a long step-time between iterations must be used to achieve a practical computation speed; this results in a sacrifice in accuracy). Also, the similarity score is influenced by isotope peaks that weigh the same as other spectral peaks and by high-intensity peaks that weigh too much. A difference in the predicted isotope pattern may significantly decrease the similarity score of the correct peptide, whereas a coincidental match of a high-intensity ion may significantly increase the score of an incorrect peptide. The challenge for applying the model to proteomic analyses is increasing the speed (currently, it is too slow for database searching with on-the-fly predictions). Potential solutions to increase the speed include using this for validating results after a database search with loose criteria, and precalculating a theoretical MS reference library. Theoretical spectra are useful as references because they are free of noise and have no incorrect peptide assignments. Also, they can be produced quickly and include spectra that contain few sequence ions. Theoretical spectra may be better than low-quality experimental spectra such as those from shotgun proteomics analyses.

Dr. Zhang concluded by noting that physical modeling is a practical method for predicting CID- and ETD-generated peptide fragmentation spectra. The CID model can be applied successfully to produce reliable peptide identification, especially for spectra with limited sequence information. The model can also be used to improve reliability of peptide identification in proteomics experiments.

Discussion

One discussant asked whether the model provided any indication as to when a prediction would not be successful. Dr. Zhang observed that the model currently needs additional optimization with prediction of neutral loss of water and similar modifications.

Spectrum Library Building and Searching

Henry Lam, Ph.D., Institute for Systems Biology

Dr. Lam began by observing that database searching is slow and can be error-prone, yet it is the current standard for proteomic bioinformatics. The approach is a "memory-less" workflow, as most data are stored locally and ultimately not shared. By contrast, spectral libraries are collections of spectra that are identified initially by sequence searching but enable future identifications by spectral matching (e.g., similar spectra are inferred to imply the same identification). Advantages of spectral libraries include speed and high sensitivity and specificity; however, identifications are limited to those proteins previously observed. The MS platform currently represents the limiting factor in spectral library approaches. Targeted approaches will answer the most interesting biological questions, and spectral libraries can be integrated with sequence searching to improve the data pipeline.

Dr. Lam then noted that a peptide identification pipeline of the future will begin with upfront peptide library comparison. Peptides not identified at this point can then be checked against libraries of unidentified peptides for possible matches, and a quality filter can help to determine which spectra are likely to be viable. Many spectral searching tools have been developed, including the NIST tool (Stein S et al. *J Am Soc Mass Spectrom* 1994), LIBQUEST (Yates J et al. *Anal Chem* 1998), X!Hunter (Craig et al. *J Proteome Res* 2006), Bibliospec (Frewen et al. *Anal Chem* 2006), and SpectraST (Lam et al. *Proteomics* 2007). An overnight SpectraST search of a human plasma dataset (2.4 million MS/MS spectra) against a NIST human spectrum library resulted in 430,000 identifications. A similar SEQUEST search against the human International Protein Index (IPI) identified 350,000 peptides over the course of a few weeks. Of the two approaches, SpectraST identified 22 percent more spectra for a given confidence cutoff limit. Extra SpectraST identifications were mostly the same identifications made by SEQUEST but at lower confidence and repeated identifications of the same peptide ions from lower quality spectra that SEQUEST failed to identify. Dr. Lam noted that SpectraST is rarely wrong with confident identifications (~0.01 percent false-discovery rate). As much as 99.5 percent of the identifications missed by SpectraST were not in the spectral library.

The improvements seen in this comparative analysis result from a smaller search space (e.g., only previously observed peptides), identifications from several search engines contained in the NIST library that implicitly combine multiple search algorithms, and more precise similarity scoring afforded by global similarity (rather than only the presence of *b*- and y-ions).

Spectral libraries can be constructed through centralized (e.g., NIST, PeptideAtlas, Global Proteome Machine Project [GPM]), and self-generated approaches (e.g., X!Hunter, SpectraST). The former option offers greater coverage and higher quality, but the latter can include proprietary data that can be organized and reused. When building a library, one must consider the quantity of spectral information to include in the spectra and how to handle multiple observations of the same peptide ion. For the latter issue, a consensus approach can be used to create a "consensus" spectrum. A scoring metric can be used to

determine which replicates should be omitted from the consensus. However, quality control (QC) measures must be built into the software to identify false positives and similar spectra that have conflicting identifications. Using the PeptideAtlas, 40 datasets can be built into a spectral library in one day.

Dr. Lam concluded that SpectraST was approximately 1,000 times faster than SEQUEST and generated up to 67 percent more confident identifications at given false-discovery rates. Moreover, the quality of the library is essential; a higher quality library affords better discrimination and more identifications at the same confidence cutoff. However, there will be an eventual tradeoff with lower coverage. Consensus approaches are improvements over those based on best replicates. Although spectral libraries offer great potential for database searching, optimization efforts are needed with respect to integration, user-friendliness, data gathering, statistical validation, and algorithm development.

Discussion

Dr. Lam noted that the discriminating ability and speed of SpectraST are based on the database against which the search occurs. However, a decreased search space greatly improves this process. Decoy spectra were created using spectra from other organisms and peak shuffling.

Assigning Confidence to a Peptide ID

Alexey Nesvizhskii, Ph.D., University of Michigan

Dr. Nesvizhskii noted that several different statistical methods have been developed in recent years to validate peptide assignments in MS/MS spectra. He noted, however, that the central issues are the same for database searching and spectral library searching—the degree of similarity between spectra is compared regardless of whether reference spectra are acquired or theoretical. Ideally, the best-scoring peptide will represent the correct identification. However, this outcome is not achieved in most cases, leading researchers to explore methods that increase their confidence in their peptide identifications.

To assess confidence in peptide identifications, commonly used methods include single-spectrum and global (whole-database) measures (e.g., expectation values, target-decoy database searching, and PeptideProphet). Each method has uses and limitations (Nesvizhskii AI et al. *Nat Methods* 2007;4:787-797). Single-spectrum measures (e.g., BLAST) involve *p*-values or expectation values, which are not well suited to large-scale analysis of datasets because they do not allow estimation of error rates as a function of filtering threshold. Moreover, expectation values do not reflect additional useful information (e.g., the number of missed cleavages, peptide retention time, mass accuracy) and require that protein probabilities be computed by combining probabilities of peptides that correspond to the same protein. Whether peptide expectation values can be used for this purpose is not clear.

Decoy databases, by contrast, fail to capture the complexity of protein modifications. The mixture modeling approach (Keller A et al. *Anal Chem* 2002;74:5383) computes peptide probabilities by unsupervised clustering. Mixture model algorithms learn the most likely distributions among correct and incorrect peptide assignments given the observed data. Flexible mixture models allow multiple sources of analysis and can be used for larger datasets. However, when using MS/MS analysis, many steps can contribute to error in peptide identification or missed peptides, including digestion, depletion, separation, MS analysis, and database searching. Confident identification depends on accuracy at each of these levels, and all information must be considered in context.

Dr. Nesvizhskii then described a strategy that combines database searching with a selected peptide separation coordinate, such as isoelectric point (pI). The combination of search score with pI value can discriminate between likely matches and therefore improve statistical power. One commonly used

approach to add pI information to the pipeline is to measure pH experimentally and calculate delta pI (the difference between measured and expected pI, as based on pH). However, this approach is difficult to carry out in practice. Instead, a data-dependent statistical model that does not require pH measurement can be used to incorporate pI information. In this approach, the data indicate the expected pI values for each fraction, and the combined database search scores and predicted pI values improve the discrimination between candidate peptide identifications. This model assumes that individual residues behave independently when contributing to the pI of a peptide.

Dr. Nesvizhskii also briefly noted that investigations are under way to couple the output from multiple search tools (e.g., Mascot, SEQUEST) to increase accuracy and sensitivity of peptide identification (Searle B et al. *J Proteome Res*, submitted). Also, MS/MS and MS/MS/MS matching statistics are being investigated to increase peptide identification confidence by coupling consecutive mass spectral stage data (Ulintz P et al. *Mol Cell Proteomics*, in press).

Discussion

In response to a participant's query, Dr. Nesvizhskii noted that his laboratory has incorporated sequence tagging into its methods, although no results are currently available for this approach. He noted that confidence with the methods is assessed using Bayesian sampling.

Combining MS/MS Sequence and Spectral Library Search Algorithms

Lewis Geer, Ph.D., National Center for Biotechnology Information, NIH

Dr. Geer began by reviewing sequence searching methods that utilize experimental data and spectral libraries. Each strategy has advantages, suggesting that a combination of the two approaches could increase sequence coverage and increase sensitivity relative to each individual technique. With sequence searching, the search space is determined by the sequence library, which is required to identify new spectra. However, this strategy allows limited use of intensity and may not identify noncanonical ion species or ions that are not present. The relatively large search space also reduces sensitivity and speed. Spectral library searching, by contrast, uses peak intensities to improve scoring sensitivity and specificity. Moreover, it offers the advantages of comprehensive annotation of ions in the library spectra and limited search space. However, limiting the search space to the spectral library will also fail to identify novel spectra. Combining these two methods can therefore prove complementary.

Issues when combining these two search methods include creating a pipeline and combining the scores. Recently, the NIST MS spectral library search and the Open Mass Spectrometry Search Algorithm (OMSSA) sequence search were combined to obtain high sensitivity and comprehensive sequence coverage. Measured spectra were searched against the NIST spectral library, and peptides were mapped to protein records. Spectra without high-scoring matches were then subjected to the OMSSA sequence search. The false-discovery rate calculation used decoy databases (e.g., fly, yeast, *D. radiodurans*). The prerelease NIST 2007 MS/MS library covers 180,000 consensus human spectra and 76,000 consensus yeast spectra (total amino acid coverage relative to the human IPI, ~12 percent; total detectable amino acid coverage, ~18 percent).

One attendee asked why these coverages were in the stated ranges, and Dr. Geer noted that the value would increase if there were targeted efforts to collect tissues and build the database. Another participant noted that false positives created by homology will be influenced by decoy choice, which could present an issue with the creation of an empirical error rate.

Dr. Geer then asked participants to consider aspects necessary for inclusion in a protein sequence library of use to the general proteomics community, given the caveat that a tradeoff must be made between

coverage and sensitivity. He cited several categories of desired peptides, including coding single nucleotide polymorphisms (SNPs), alternative splicing, ab initio prediction, mature peptides and signal peptides, and noncanonical start/stop codons. He noted also that changes in library or database version must follow a clear set of criteria related to changes in annotation or sequence. He suggested that a focus on elements presently available could prove useful before attempting to expand resources. He noted, however, that varying protein preparation methods (e.g., digestion parameters) will alter the peptides that are identified. In the ensuing discussion, participants suggested considering msIPI (Schandorff S et al. *Nat Methods* 2007;4:465-466) or establishing a standard with known mutations and modifications.

User Community Session

Perspective and Progress in Protein and Peptide Identification

Sean Seymour, Ph.D., Applied Biosystems

Dr. Seymour described the Paragon™ algorithm (Shilov IV et al. *Mol Cell Proteomics* 2007;6:1638-1655) that rates short tags in terms of their relative quality, using continuous measures and preserving information as long as possible. The strategy is based on the premise that, for the search of a single spectrum, the allowed search space applied to each sequence region should be proportional to the degree of tag evidence implicating that region. Sequence "temperature values" can be computed to indicate the degree to which that portion of the database is likely to contain tags of interest. The strategy assumes that it is worth more effort to find a molecular weight match for sequence regions that contain an abundance of tag evidence. This algorithm applies different search space sizes to different parts of a database. For segments in the database that are deemed "hot," all modifications, digestion features, and tolerances are searched; only the most likely modifications are considered for "cold" segments. This approach utilizes feature probabilities to generate a continuous description of search space elements. This algorithm also enables a new kind of user interface through which users enter digestion, alkylation agents, and other parameters. Recalibration based on these entries then occurs. Peptide identification ambiguity is preserved in this process; it is not sufficient simply to track all the proteins that share a single peptide hypothesis. Preserving protein ambiguity is a key element to enable comparative analyses.

Dr. Seymour then described several examples that utilized Paragon™ to assess the relative quality of protein inference, to measure error rates, and to carry out comparative analyses. He recommended that software developers and users agree to report results of analyses at critical error rate values. He concluded by noting that data format standards are needed and that funding agencies can accelerate their implementation. The community should design a timeline toward a requirement of the submission of intermediate data and results using standard data formats, ontologies, and minimal content. He also stressed the importance of cooperation among granting agencies and established resources that share common interests. He noted that from the vendor's perspective, a single standard and fewer community efforts will be supported and implemented more quickly and effectively than will a multitude of redundant or overlapping efforts.

Discussion

One attendee asked about scoring using the Paragon™ algorithm, and Dr. Seymour noted that scoring is based on the empirical dataset and is not calculated for the dataset being analyzed. Another discussant asked how the algorithm compares to ProteinProphet, and he noted that the two agreed closely on a recent cross-comparison.

Getting Your Software Out There

Brian Searle, Proteome Software Inc.

Mr. Searle noted that commercial search engines are numerous, and some are isolated to specific instrument types. Moreover, search engines designed by academic developers proliferate. These diverse tools are often innovative, highly adaptable, and transparent in their application. However, they can also be awkward, as they may contain numerous software bugs, be difficult to use, and be saddled with intellectual property (IP) issues. With regard to IP, Mr. Searle noted that patents are prohibitive, and that academic institutions own the IP for tools developed by their employees. Furthermore, the patenting of software is time-consuming and may hinder its ultimate use. Thus, the only way for software developers to protect themselves should they wish to take the application upon leaving the university is to use open-source (OS) code. He noted also that graphical interfaces and ease of use (e.g., "fixing bugs") are critical.

Since launching Scaffold in June 2005, Proteome Software Inc. has discovered more than 1,000 bugs in the software (13 percent in algorithms, 23 percent in file loading, and 64 percent in the graphical interface). Thus, continued software support is primarily user-interface work. Mr. Searle noted that as the distribution becomes broad, costs increase. However, use escalates as well, thereby decreasing the unit cost/user. Nonetheless, he noted that none of this work results in academic publications, and lack of publications means that investigators must have an interest in the software that goes beyond the short term. He also advocated hiring professional software engineers with the recognition that very little funding is available for system maintenance. He suggested that funding of the caliber of R01 awards is not necessary for software development support; instead, award amounts such as those provided through Small Business Innovation Research (SBIR) awards (e.g., $100,000) should be sufficient. However, the iterative design phases of software development and refinement are not amenable to SBIR and R01 funding strategies. Therefore, Mr. Searle suggested that smaller grants with funding tied to an iterative design framework will support the development of software.

Discussion

Participants discussed the issue of funding for software development, noting that there is currently no R01 that is dedicated to software development. Rather, R01 support is given on the condition that software is proven usable. Another participant noted that software is often developed in academic laboratories to support specific research goals, and these applications should also be funded. One discussant asked how a company could use OS code and still be commercially viable. Mr. Searle noted that OS software can be used as the basis for and be bundled into a commercially viable application.

Applications for Open-Source Software in Proteomics

Philip Andrews, Ph.D., University of Michigan

Dr. Andrews noted that open source (OS) software is important to support technology development. Given that instrumentation and methodologies change rapidly, OS software enables the development and dissemination of new and improved applications. He noted that software development always lags behind hardware and applications. Since proteomics features large, complex datasets, the field requires significant investment in its computational infrastructure. For these reasons, nearly all technology development in proteomics requires the development or modification of computational tools. However, instrument development is often hampered or negatively impacted by lagging software development. Yet if the market for software is perceived to be small, resources will not be allocated to support it. Partial solutions to this bottleneck include allocating additional resources toward proteomics informatics (on the order of one-third of a proteomics budget) and building and supporting an OS community.

Open source is defined by WordNet as “of or relating to or being computer software for which source code is freely available.” Dr. Andrews observed that this definition does not mean that the software is necessarily free. However, OS software is not hampered by licensing issues that depend on various policies for the developer, home institution, or funding agencies or on existing IP arrangements. Instead, OS software can be licensed under Creative Commons licenses that stipulate use, modification, or distribution of code with proper attribution. OS is useful for proteomics because it allows interpretation of data from new technologies, enables new approaches to proteomics, enables multiple uses of datasets, and provides documentation for new algorithms and tool sets for developers and researchers. Applications of OS software thus represent useful examples of protocols.

Dr. Andrews noted also that synergies can exist between OS and commercial tools; there is a degree of progression from OS tools to commercial packages as a field matures. OS tools can thus inform commercial software development, either directly or indirectly. Some OS software can be used directly by commercial platforms (e.g., X!Tandem/Scaffold), while other OS software will ultimately make the transition to commercial products.

In summary, Dr. Andrews noted that a robust OS community will accelerate progress in proteomics. OS efforts allow the community to respond rapidly to new technologies and to emerging needs. The cost of development is shared, and OS is compatible with commercial proprietary software. With respect to proteomics, OS code must be publicly available and free to use; technically, it can be considered as “free software.”

Dr. Andrews then provided perspective on open access to data, which can be defined as free and public access to digital scholarly and scientific information. Open access has been enabled by the Internet (e.g., PubMed Central, Public Library of Science, Creative Commons). While the traditional print article format has remained constant for 300 years, proteomics is a data-centric field. There is at present a disjunction between printed scholarly articles and contemporary data generation and information transfer. Larger, more complicated data are hard to reduce to print form and to review properly. As such, potential data loss occurs over time, as researchers move and institutions change their data-storage policies. Reproduction of results and the extension of public knowledge then lead to an increasing marginalization of the original purpose of the data.

Thus, open access to datasets is useful for proteomics, since datasets are required to validate algorithms and software. Moreover, re-use of datasets informs genomic annotation efforts, biomarker discovery, and systems biology initiatives and validates data interpretations in all of these applications. Several open-access proteomics databases currently exist (e.g., PeptideAtlas, GPMDB, PRIDE, Human Protein Reference Database [HPRD]). Dr. Andrews also discussed Tranche, a secure data repository and transactional system built on a peer-to-peer data-sharing model. Tranche offers secure dissemination of public or private data with proper citation, data provenance, and integrity. Tranche use is independent of file format, and the system allows forward annotation, file retraction, and related capabilities. The application currently has more than 5,000 users and is linked to PRIDE and HPRD.

Next Steps and Closing Remarks

Dr. Kinsinger thanked speakers and attendees for their candid and thought-provoking discussions. He noted that NCI and NIST are seeking ideas and ways that the agencies can help move the field forward. He also stated that many goals thought to be impossible several years ago are coming to fruition with respect to bioinformatics and proteomics. He thanked participants once again for a productive meeting, and the meeting was adjourned.