PROTEOTYPING OF MICROBIAL COMMUNITIES BY OPTIMIZATION OF TANDEM MASS SPECTROMETRY DATA INTERPRETATION

ALYS HUGO[1], DOUGLAS J. BAXTER[2], WILLIAM R. CANNON[1], ANANTH KALYANARAMAN[4], GAURAV KULKARNI[4], STEPHEN J. CALLISTER[3]

[1]*Computational Biology and Bioinformatics Group,*
[2]*Molecular Sciences Computing Facility,*
[3]*Biological Separations Group,*
Pacific Northwest National Laboratory
Richland, WA 99352

[4]*School of Electrical Engineering and Computer Science,*
Washington State University,
Pullman, WA, 99164.

We report the development of a novel high performance computing method for the identification of proteins from unknown (environmental) samples. The method uses computational optimization to provide an effective way to control the false discovery rate for environmental samples and complements *de novo* peptide sequencing. Furthermore, the method provides information based on the expressed protein in a microbial community, and thus complements DNA-based identification methods. Testing on blind samples demonstrates that the method provides 79-95% overlap with analogous results from searches involving only the correct genomes. We provide scaling and performance evaluations for the software that demonstrate the ability to carry out large-scale optimizations on 1258 genomes containing 4.2M proteins.

1. Introduction

Peptide and protein identification from global proteomics studies of environmental samples is challenging because the standard approach used to identify peptides and proteins is difficult to effectively apply. The standard approach involves growing clonal cells in the laboratory, separating the proteins from other cellular components, digesting the proteins into smaller peptides (approximately 6-55 amino acids long), partially separating the peptides using liquid chromatography, and then introducing each peptide into a mass spectrometer.

The peptides are then collided with an inert gas that vibrationally excites the peptide. The vibrational excitation leads to fragmentation of the peptide into smaller chemical components, which are often short sequences of amino acid residues. The most widely used method to identify the peptide from its mass spectrum is to compare a predicted fragmentation pattern of the peptide (a model spectrum) to the experimental spectrum. This approach involves a search of all possible peptides that can be generated from the protein sequences. There are a number of available analysis tools that perform these searches, including Sequest [1], X!Tandem [2], Crux [3], Mascot [4], OLAV [5], InsPecT [6], and MSPolygraph [7-9].

Because the model spectra used in these searches are typically developed by training over diverse sets of peptide spectra, the resulting model spectra are often not good representations of the true spectrum, which can result in low sensitivity [9]. Given the high mass accuracy of many spectrometers, this would not be a problem except for the large number of peptides that can be generated from the protein sequences of a single microorganism, resulting in a high chance of a

random match. This, of course, reduces the number of peptides that can be confidently identified. For instance, when controlling the false discovery rate at 5%, often only 15-20% of the spectra can be identified with a peptide.

In an environmental sample, the problem is exacerbated because there may be several to hundreds of different microbes present. The vast number and diversity of possible peptides that have to be searched leads to a very low number of spectra that can be confidently identified at a reasonable false discovery rate.

Approaches to identify peptides and proteins from environmental samples have generally taken one of two approaches for searching protein sequences. One approach is to search against all known protein sequences using sequence databases such as the NR database at NCBI or likewise the Swiss-Prot database at the Swiss Institute of Bioinformatics. However, this approach has not been favored because the large number of protein sequences leads to a high chance of a match simply by chance. Alternatively, recent environmental proteomics studies have sought to limit the number of potential candidate peptides by searching against metagenome sequences from the same environment [10-12]. This is a sensible approach, but it too has many challenges. First, it requires having metagenome sequences to begin with. If one is available, protein coding regions must be determined from the draft sequences, or spectra can be matched to all potential open reading frames from the draft sequences. However, the problem with the latter approach is similar to searching against large databases such as NR or Swiss-Prot: the chance of a random match is quite high due to the large number of candidates and the moderate specificity of the model spectra that are used for identification.

Frequently, metaproteomics studies also supplement the analysis with *de novo* peptide identification, which seeks to determine the identity of the peptide responsible for the spectrum only from the peaks in the spectrum. Unfortunately, peptide MS/MS spectra most often do not provide enough information to obtain full-length peptide sequences and are challenging to apply to spectra from peptides with charges greater than +2.

As an alternative, access to high performance computing resources allowed us to developed an optimization method for identifying peptides and proteins from fully sequenced microbial genomes. The method searches all fully sequenced genomes and optimizes proteome-spectra matches by iteratively eliminating microbes that are not likely to be in the sample. The method has been tested using samples containing blind mixtures of spectra from known microbes and samples containing unknown mixtures of microbes. In the case of five blind mixtures of varying complexity, the method has been able to identify the correct microbes reliably. In addition, the spectra identified with each microbe has a high overlap with spectra identified at a 5% FDR when searching only the protein sequences of the correct organisms.

2. Methods

The method evaluates which known genomes are appropriate for analysis of an environmental sample using the statistics based on the number of top hits observed for each genome in the MS/MS data. Specifically, a database search tool, *MSPolygraph* in this case, is used to evaluate matches between each spectrum and candidate peptides from the protein sequences of each of the fully sequenced genomes available. In the work here, we used the 1258 fully sequenced bacterial and fungal genomes available from NCBI on Oct 4, 2011. Each candidate peptide may correspond to one or more genomes, and genomes can have multiple candidate peptides for each spectrum. However, for a particular spectrum only the highest scoring peptide from a genome is

recorded. If the genome has no candidates for the spectrum, the genome's score for the spectrum is zero. Spectra with no matches are removed. The genome having the peptide with the highest score for a spectrum is referred to as the top hit for that spectrum. Also, for each genome, the total number of candidate peptides for that genome in each spectrum is recorded. The significance of each genome is then calculated based on a statistical likelihood. A large number of ties occur in this process, in which a peptide that is a top hit to a spectrum belongs to proteins from multiple genomes. The ties are broken in an iterative manner with an expectation-maximization approach. Details on the method follow.

2.1 *Statistical Null Model*

The probability that a particular genome would appear as the top hit for a particular spectrum by chance is estimated as the number of peptide candidates generated from the protein sequences of the genome divided by the total number of candidates from all genomes for that spectrum. That is, the probability of a random top hit for a particular genome to a spectrum is proportional to the relative number of peptide candidates that can be derived from the protein sequences of that genome. From these probabilities, we estimate the number of top hits that would be expected to be observed for each genome. This is the number of spectra multiplied by the null probability. In addition, the information on the null probabilities is later used to generate a population of top hits that a genome would obtain by chance for each spectrum from a simulation.

2.2 *Sample likelihood ratio and significance*

For each genome, we calculate the sample likelihood ratio, which is the likelihood of observing the genome from the data relative to the likelihood of observing the genome by chance. Accordingly, the sample likelihood is calculated as the number of top hits observed from the data $n_{obs,i}$ relative to the number of top hits expected from the null probabilities $n_{null,i}$,

$$\theta_i = \frac{n_{obs,i}}{n_{null,i}}.$$

To estimate the significance of the sample likelihood ratios, the number of top hits expected by chance for each genome is estimated. As mentioned above, this is the number of spectra multiplied by the null probability. Next, using the null probabilities, a simulation is then used to generate a population of likelihood ratio values that would be expected to be observed by chance. The population of likelihood ratios is then fit to a generalized extreme value distribution. From the fit, the probability of observing the sample likelihood ratio by chance (p-value) is estimated.

2.3 Iterative assessment of likelihood

Initially, distribution of likelihood scores is obtained in which no particular genome stands out, in large part because many top-scoring peptides are shared among them. In this case, it is not clear which genomes are good proxies for the organisms contained in the environmental samples. Ideally, ties are broken by assigning each involved genome a fraction of the top hits proportional to the their abundance in the environmental sample. However, the proportion of each species in the sample is unknown, and the number of observed top hits is a conflation of hits due to the microbe actually being present and random hits to the genome of the microbe. Instead, ties are broken by assigning each genome a fraction of the top hits proportional to the sample likelihood ratio for the genome determined in the previous iteration and scaled by the probability that the sample likelihood ratio was not by chance (1- p-value). Specifically, the quantity C_i, as calculated below, is found for each top scoring genome i and normalized over all top scoring genomes for the spectrum, resulting in each genome's share of the top hit,

$$C_i = \theta_i * (1 - p_i).$$

Here θ_i is the sample likelihood ratio for genome i from previous iteration, and p_i is the p-value of the sample likelihood ratio for genome i, calculated using the results of simulation. New θ and p-values are calculated by totaling the fraction of top hits assigned to each genome and comparing the results to simulated results.

This process is repeated for a predetermined number of iterations or until some convergence criterion is met. (In our case, 150 iterations or until top 50 θ values changed by less than 0.0001.) After the last iteration, ties are reassigned if desired. For each spectrum, the tied genomes with the most top hits overall from the last iteration receives full credit for the hit. This gives a simplified explanation in which the spectra are accounted for with as few genomes as possible. In most cases, this results in consolidation of hits from several closely related genomes (often species from the same genus), leaving one representative that best explains the observed hits. This step is most useful when one expects only single representatives of each genus to be present and can be omitted if multiple species from a genus are expected to be present. *P*-values are calculated for each genome based on the recalculated number of hits using results from a second simulation.

2.4 False Discovery Rate Estimation

Finally, genomes are determined to be appropriate or not based on a desired false discovery rate. The false discovery rate is estimated from the final set of *p*-values using a nonparametric estimation with the program *qvality* [13]. If a genome has posterior error probability less than 1%, it is considered to be appropriate for use in the analysis. If a genome fails to meet this requirement, it is considered to be inappropriate.

2.5 ***Programming Model***

The statistical methods discussed herein were implemented in Matlab, and run on common workstations. The database search tool *MSPolygraph* had to be modified to handle protein sequences from thousands of genomes and to keep a list of only the top scoring peptides from each genome. As the number of genomes is constantly growing, high performance computing is required.

The parallel version of the *MSPolygraph* is essentially a task-scheduling wrapper around the serial version of the code. In the serial code an input parameter file is read, along with the fragmentation model for generating model spectra, the protein sequences of all organisms to be searched, and the spectra file(s) to be analyzed. The code loops over the spectra scoring each spectrum against all peptides generated from the target protein sequences that are consistent with the observed mass-to-charge ratio of the observed, intact peptide, accumulating high scoring matches and printing them out in a sorted list, sorted by likelihood ratio score. The amount of work required to score a spectrum depends on a variety of factors which include its length (number of peaks), how many peptide candidates were generated from the target protein sequences, and how many peaks from each of those peptide candidates match peaks from the experimental spectrum peaks. The number of peaks from each peptide that will match the spectrum cannot be predetermined without doing 2/3 of the work required to actually score the spectrum, which makes an *a priori* determination of run time for a spectrum impractical. Hence, we use a dynamic scheduling scheme facilitated by a server/client process model.

In the parallel version we use the MPI (Message Passing Interface) standard for communication. The input files are placed on a globally visible file system (mounted on all of the compute nodes). Each processor reads in to its own local memory the input files and we then use a dynamically scheduled server-client model to control which process (mpi rank) scores which spectrum. A processor's behavior is controlled by its mpi rank. One processor is the dedicated server process (mpi rank 0) and all others are considered client processors.

After reading the input data, the server process issues a non-blocking receive to each client. It uses a simple counter to determine which task to send to a requesting client. It polls clients for responses indicating that a spectrum has been completed (or during the first pass that a client has initialized and is ready to start scoring), and replies with another index for a spectrum to be scored if there is one or a quit message if all tasks have been distributed. The manager utilizes non-blocking sends so as not to need to wait for the clients to receive their messages. A different buffer for each client is used. As the outgoing messages are only an index as to which spectra to score, and the incoming messages are a fixed length summary line, only a small amount of space is required even for a very large number of clients. Also, as the server hands out a new index to a "ready to start" or "completed spectrum" message from a client, no more than 1 message per client is ever in flight. After all tasks have been handed out, the server processor continues polling for responses till responses for all spectra have been received.

The client process also reads the input files from the global file system, opens its own output file for printing scoring results for spectra assigned to it. It then issues a non-blocking receive for an index of the next spectra to be score and sends a "ready to start" message to the server processor. The client then enters a communication and scoring loop in which the client repeatedly:

- waits for server message indicating which spectrum to process (or a quit message),
- processes the spectrum (or exits the loop),
- writes data and flushes the results to its output file,
- issues a non-blocking receive for the next spectrum index and,
- sends a summary message for the scored spectrum to the server.

Scaling results are presented below.

3. Results

3.1 Optimization of Artificial Microbial Communities

Five laboratory samples were prepared of varying complexity, ranging from 3 to 15 microbes. The samples were mixed in equal proportions based on estimates of total protein concentration. The results for the lowest complexity sample are shown in Table I. In this sample, three species were present, *Anaeromyxobacter dehalogenans, Geobacter uraniumreducens* and *Salmonella typhimurium*. As shown in Table I, these three species were identified as the top three species by the optimization procedure and were the only species identified at a 1% false discovery rate (q-value). For *Salmonella typhimurium*, however, a strain that is 98% identical to the cultured strain, *Salmonella typhimurium LT2,* was identified (*Salmonella enterica serovar Typhimurium*). For each of these species, the number of spectra associated with the species is shown in the second column of Table I. For comparison, the number of spectra found for each of the species using a standard identification procedure is also shown in the third and fourth columns of Table I. In the standard approach, only the proteomes of the known species are searched, and the number of spectra identified with each of the species at 5% and 10% false discovery rates is shown in these latter columns. Also shown in parentheses in columns three and four are the percent of peptide-spectrum matches occurring in the optimization approach that also are found

Species	Number of Identifications				
Target Species Other Species	Optimization	Overlap 5% FDR	Overlap 10% FDR	LR	Posterior Error Prob
Salmonella enterica serovar Typhimurium	3145	3016 96%	3081 98%	289.20	0.03
Anaeromyxobacter dehalogenans 2CP-C	1256	1199 94%	1249 97%	67.12	0.06
Geobacter uraniumreducens Rf4	539	525 95%	533 96%	39.83	0.08
Below 1% q-value cutoff:					
Salmonella typhimurium LT2	6	0 0%	8 38%	0.57	1

Table I. Optimization results for model microbial community containing three known microbes.

in the standard approach. For this dataset, the overlap in peptide-spectrum matches between the optimization method and the standard search is remarkably high, 94-98%. That is, despite searching 1258 genomes containing over 4 million proteins and several orders of magnitude more peptides, the optimization approach on this data set contained nearly the same quality of identifications as a standard database search of only the proteins correct genomes. Of course, as a model of a microbial community found in the environment, three organisms would be a very simple microbial community.

To test the ability of the method to analyze more complex samples, we also investigated four other datasets consisting of 6, 9, 12 and 15 microbes. The results for the highest complexity data set, derived from the mixture of batch cultures from 15 species, are shown in Table II. For this data set, each of the top 10 scoring species and 11 out of the top 12 scoring species were correctly identified. In all, 12 of the 15 species were identified at an estimated 1% false discovery rate. For these 12 species, the overlap in peptide-spectrum matches between the optimization method and the standard approach (identification using only the correct genomes) was 47-91%. For all 15 species, the weighted average of the overlap was a remarkable 79%. While this is less overlap than was seen in the data containing just three microbes discussed above (95% weighted average overlap), the performance was encouraging when the reasons for the decreases were examined. These are discussed are discussed next.

Three species were not identified - *Rhodopseudomonas Palustris CGA009, Saccharomyces cervisiae S288c,* and *Aspergillus carbonarius*. For *Aspergillus carbonarius*, 62 peptide-spectrum matches were found for a species from the same genus, *Aspergillus niger*. It is possible that determining the likelihood at the genus level instead of at the genome level would identify the *Aspergillus* genus. However, the number of significant peptide-spectrum matches found for each of these species was also quite low when a standard approach (only the correct protein sequences) was used. For instance, for *Aspergillus carbonarius* only 49 and 103 significant peptide-spectrum matches were found at 5% and 10% false discovery rates. The most likely cause is that the protein samples that were prepared for these species were too dilute. For *Aspergillus carbonarius*, a confounding problem is that the protein sequences are only available from the initial draft genome of the organism.

Species	Number of Identifications				
Target Species Related Species Other Species	Optimization	Standard run on known DBs			
		5% FDR	10% FDR	LR	Posterior Error Probability
Deinococcus radiodurans	1168	1048 83%	1191 90%	165.92	1.79E-05
Shewanella oneidensis	713	627 85%	714 91%	76.94	1.32E-04
Salmonella typhimurium LT2	482	495 83%	554 90%	50.86	3.86E-04
Synechococcus PCC 7002	247	230 84%	268 90%	41.70	6.47E-04
Arthrobacter FB24	323	299 85%	347 89%	31.58	1.33E-03
Heliobacterium modesticaldum Ice1	239	225 79%	269 87%	30.74	1.43E-03
Chloroflexus aurantiacus J 10 fl	226	208 80%	259 89%	26.81	2.04E-03
Desulfovibrio desulfuricans G20	177	153 77%	188 86%	20.19	4.27E-03
Anaeromyxobacter dehalogenans 2CP-C	265	243 66%	325 76%	15.77	8.11E-03
Geobacter uraniumreducens Rf	92	91 70%	119 79%	7.37	5.83E-02
Brucella melitensis bv	39			4.99	1.59E-01
Candidatus Korarchaeum cryptofilum	21			4.50	2.08E-01
Clostridium thermocellum ATCC 27405	40	41 67%	54 76%	4.03	2.77E-01
Beutenbergia cavernae DSM 12333	33			3.50	3.97E-01
Rhodobacter sphaeroides KD131	38	46 47%	79 52%	3.49	3.99E-01
Burkholderia cenocepacia J2315	59			3.06	5.60E-01
Acidovorax avenae citrulli AAC00-1	38			3.02	5.78E-01
Frankia alni ACN14a	53			3.00	5.91E-01
Desulfurivibrio alkaliphilus AHT2	21			2.88	6.51E-01
Rhodococcus erythropolis PR4	36			2.82	6.83E-01
Xylella fastidiosa	15			2.79	7.06E-01
Below 1% q-value cutoff:					
Rhodopseudomonas palustris BisB18	27			2.05	1.00E+00
Aspergillus niger CBS 513 88	62			1.31	1.00E+00
Rhodopseudomonas palustris BisB5	14			1.18	1.00E+00
Rhodopseudomonas palustris BisA53	12			0.91	1.00E+00
Rhodopseudomonas Palustris CGA009	10	23 5%	43 10%	0.79	1.00E+00
Saccharomyces cervisiae S288c	15	17 21%	42 36%	0.61	1.00E+00
Aspergillus carbonarius	18	49 7%	103 25%	0.48	1.00E+00

Table II. Optimization results for model microbial community containing 15 known microbes.

Likewise, *Rhodopseudomonas Palustris* CGA009 was also not identified at a statistically significant level. In this case, 10 peptide-spectrum matches were found in the optimization, while 23 and 43 matches were found when using a standard search of the known genomes only. This indicates that the sample was relatively dilute in peptides from this organism, but interestingly 43 peptide-spectrum matches were also assigned to closely related *Rhodopseudomonas* species. In total, 53 matches were found to the Rhodopseudomonas species, which is more than found in the

standard search. This may indicate that the species grown in culture is no longer clonal with respect to the sequenced species *Rhodopseudomonas Palustris* CGA009. In this regard, the optimization approach may be useful for analyzing batch monocultures also against closely related genomes and would effectively allow for sequence variability in the peptide-spectrum matching process.

The third species that was not identified was *Saccharomyces cervisiae* S288c. In this case, 15 peptide-spectrum matches were found in the optimization, while only 17 and 42 matches were found at 5% and 10% FDR in a standard search of the known genome. However, even if the same number of matches were found in the optimization as was found at 10% FDR in the standard search against the known genome, this would not have been enough matches to be statistically significant. Again, the most likely explanation for this was that the sample simply did not contain enough proteins from this organism.

The three other data sets varied in the number of species in the sample from 6 to 12. The trends in these three data sets mirrored that seen for the other two data sets. In two of the data sets, the related species *Salmonella enterica serovar Typhimurium* was identified instead of *Salmonella typhimurium LT2*. *Aspergillus carbonarius* was consistently missed in each data set while the related *Aspergillus niger* was positively identified in one of the three data sets. While *Aspergillus niger* was not identified in the other two data sets, it did consistently have more peptide-spectrum matches (> 4-fold) than *Aspergillus carbonarius*. It is reasonable to expect that the number of *Aspergillus* peptide-spectrum matches could be significant at the genus level.

3.2 Scaling

As the processing time for a single processor job, T(1), takes longer to run on a single node than the job policy allowed at the time of these runs, we instead generate a weak scaling curve replacing T(1) by 128*T(128) shown in Figure 1. This amounts to taking the efficiency at 128 processors to be 1 for comparison purposes. We note that T1/(128*T(128)) < 127/128 (0.992), as the master processor does no work. The fall off in efficiency shown in Figure 2 at 1024 processors is an indicator that we are hitting inefficiency in the MPI infrastructure layer at scale, most likely due to too many messages being passed. This could also be also due to input/output bottlenecks if many processors are simultaneously writing output files. However, runtime

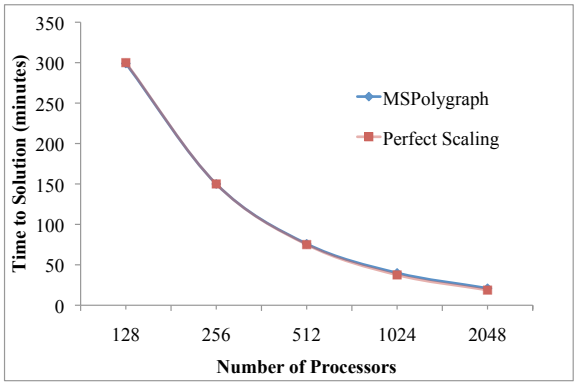

Figure 1. Scaling of MSPolygraph. The time to solution was determined for running 1258 fully sequenced genomes against 18,929 spectra on the Chinook supercomputer at EMSL.

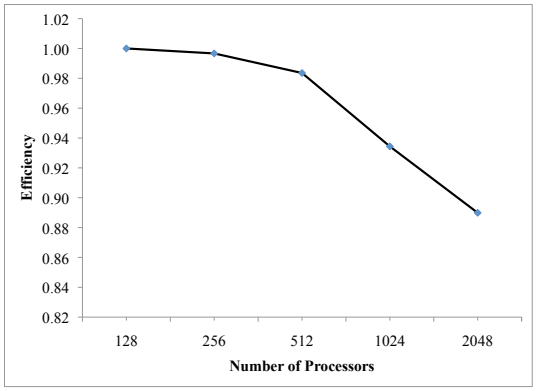

Figure 2. Parallel efficiency of MSPolygraph on data in Figure 1.

monitoring doesn't indicate that the code is any where close to the I/O bandwidth limits on the machine, making it unlikely that an I/O bottleneck causes the efficiency loss.

4. Discussion

The method discussed here allows us to use high performance computing to potential identify proteins from complex environmental samples without necessarily having sequenced the microbes in the sample. Despite analyzing over 1000 genomes against a spectral dataset, the method resulted in a 79-95% overlap with identifications made using only the correct protein sequences. There are several caveats to this approach that must be kept in mind. First, a high scoring genome does not necessarily imply that the microbe is present, only that the genome of the microbe is appropriate for analyzing the data. That being said, the presence of multiple high scoring genomes from the same genus does imply the presence of a microbe from that genus. Second, the method will actually become more powerful as more genomes become sequenced. In this regard, projects such as the Genomic Encyclopedia of Bacteria and Archaea [14], which seek to provide greater breadth of our knowledge of microbial genomes, are invaluable. At this time, there are approximately 1500 fully sequenced genomes of microbes available, and this number is increasing rapidly. At the same time, advances in instrumentation are resulting in a rapid increase in the number of spectra that are derived from a single sample. Currently, one sample may result in 10-30,000 spectra, but that number will continue to grow rapidly as the instrumentation technology develops. In summary, the statistical method described above combined with high performance computing offers a potentially significant break-through in the analysis of environmental samples.

Acknowledgements

This work was supported under contracts 57271 and 54976 from the Department of Energy's Office of Advanced Scientific Computing Research (OASCR) and Office of Biological and Environmental Research (BER) to develop new approaches for computational biology in areas of national interests. We would like to thank researchers in the laboratory of Richard Smith at PNNL who generated the datasets herein, as well as Dr. Smith, for providing the data. The calculations were performed on the Molecular Science Computing supercomputer, Chinook in EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory and at the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

References

[1] K. Eng, *et al.*, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the American Society of Mass Spectrometry,* vol. 5, pp. 976-989, 1994.

[2] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra," *Bioinformatics,* vol. 20, pp. 1466-7, Jun 12 2004.

[3] C. Y. Park, *et al.*, "Rapid and accurate peptide identification from tandem mass spectra," *J Proteome Res,* vol. 7, pp. 3022-7, Jul 2008.

[4] D. N. Perkins, *et al.*, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis,* vol. 20, pp. 3551-3567, DEC 1999.

[5] J. Colinge, *et al.*, "OLAV: towards high-throughput tandem mass spectrometry data identification," *Proteomics,* vol. 3, pp. 1454-63, Aug 2003.

[6] S. Tanner, *et al.*, "InsPecT: identification of posttranslationally modified peptides from tandem mass spectra," *Anal Chem,* vol. 77, pp. 4626-39, Jul 15 2005.

[7] W. R. Cannon, *et al.*, "Comparison of probability and likelihood models for peptide identification from tandem mass spectrometry data," *J Proteome Res,* vol. 4, pp. 1687-98, Sep-Oct 2005.
[8] W. R. Cannon and M. M. Rawlins, "Physicochemical/Thermodynamic Framework for the Interpretation of Peptide Tandem Mass Spectra," *Journal of Physical Chemistry C,* vol. 114, pp. 5360-5366, Apr 1 2010.
[9] W. R. Cannon, *et al.*, "Large Improvements in MS/MS-Based Peptide Identification Rates using a Hybrid Analysis," *J Proteome Res,* Mar 30 2011.
[10] J. F. Banfield, *et al.*, "Proteogenomic approaches for the molecular characterization of natural microbial communities," *OMICS,* vol. 9, pp. 301-33, Winter 2005.
[11] S. M. Sowell, *et al.*, "Proteomic analysis of stationary phase in the marine bacterium "Candidatus Pelagibacter ubique"," *Appl Environ Microbiol,* vol. 74, pp. 4091-100, Jul 2008.
[12] K. E. Burnum, *et al.*, "Proteome insights into the symbiotic relationship between a captive colony of Nasutitermes corniger and its hindgut microbiome," *ISME J,* vol. 5, pp. 161-4, Jan 2011.
[13] L. Kall, *et al.*, "QVALITY: non-parametric estimation of q-values and posterior error probabilities," *Bioinformatics,* vol. 25, pp. 964-6, Apr 1 2009.
[14] D. Wu, *et al.*, "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea," *Nature,* vol. 462, pp. 1056-60, Dec 24 2009.