Proteogenomics: needs and roles to be filled by proteomics in genome annotation

Charles Ansong, Samuel O. Purvine, Joshua N. Adkins, Mary S. Lipton and Richard D. Smith

Advance Access publication date 10 March 2008

Abstract

While genome sequencing efforts reveal the basic building blocks of life, a genome sequence alone is insufficient for elucidating biological function. Genome annotation—the process of identifying genes and assigning function to each gene in a genome sequence—provides the means to elucidate biological function from sequence. Current state-of-the-art high-throughput genome annotation uses a combination of comparative (sequence similarity data) and non-comparative (*ab initio* gene prediction algorithms) methods to identify protein-coding genes in genome sequences. Because approaches used to validate the presence of predicted protein-coding genes are typically based on expressed RNA sequences, they cannot independently and unequivocally determine whether a predicted protein-coding gene is translated into a protein. With the ability to directly measure peptides arising from expressed proteins, high-throughput liquid chromatography-tandem mass spectrometry-based proteomics approaches can be used to verify coding regions of a genomic sequence. Here, we highlight several ways in which high-throughput tandem mass spectrometry-based proteomics can improve the quality of genome annotations and suggest that it could be efficiently applied during the gene calling process so that the improvements are propagated through the subsequent functional annotation process.

Keywords: *proteogenomics; genome annotation; proteomics; mass spectrometry*

INTRODUCTION

Over the past decade and half, the process of completing a genome sequence has transitioned from being a challenging endeavour to being a relatively routine process for both microbial and eukaryotic species. The first complete genome sequence from a free-living organism (*Haemophilus influenzae Rd.*) was reported in 1995 [1], and was rapidly followed by sequences for other microbial genomes (*Mycoplasma genitalium, Mycobacterium tuberculosis*) [2, 3]. Since then, numerous archeal (~41), bacterial (~468) and eukaryotic (~49) genomes, including those of model organisms such as *Saccharomyces cerevisiae* (yeast) [4], *Caenorhabditis. elegans* (worm) [5], *Drosophila melanogaster* (fruit fly) [6], *Arabidopisis thaliana* (plant) [7], mouse [8], rat [9] and sea urchin [10] have been sequenced. Sequencing of the human genome was auspiciously completed in 2004, laying a foundation for advances in medical and biological research that would benefit human health [11–13]. Today, sequencing efforts continue for a myriad of other genomes of biological interest.

The completion of a genome sequencing effort represents a milestone for understanding the

Corresponding author. Richard D. Smith, Biological Sciences Division, Pacific Northwest National Laboratory, P.O. Box 999/K8-98, Richland, WA 99352, USA. Tel: +1-509-371-6576; Fax: +1-509-371-6564; E-mail: rds@pnl.gov

Charles Ansong is a Postdoctoral Research Associate in the Biological Systems Analysis and Mass Spectrometry Group within the Biological Sciences Division at Pacific Northwest National Laboratory (PNNL) in Richland, Washington.

Samuel O. Purvine is a Senior Research Scientist within the Environmental Molecular Sciences Laboratory at PNNL.

Joshua N. Adkins is a Senior Research Scientist in the Biological Systems Analysis and Mass Spectrometry Group within the Biological Sciences Division at PNNL.

Mary S. Lipton is a Staff Scientist in the Biological Systems Analysis and Mass Spectrometry Group within the Biological Sciences Division at PNNL.

Richard D. Smith is a Battelle Fellow and Technical Group Leader for Biological Systems Analysis and Mass Spectrometry within the Biological Sciences Division at PNNL.

genetic blueprint of any particular organism. However, to realize the full biological value of the sequenced genome requires accurate identification of the protein-coding genes in each genome, as well as the nature of the functional protein products. This understanding has shifted attention from genome sequencing to genome annotation. Mass spectrometry (MS)-based proteomics approaches [14–16] directly measure peptides arising from expressed proteins, which allows for direct verification of coding regions of a genomic sequence. In addition to verifying protein-coding genes, these analyses can also benefit genome annotation by helping to identify missed protein-coding genes, confirm the expression of alternative splice variants in eukaryotic genomes, and correct overestimated protein-coding potentials, particularly in microbial genomes, for example. As such, proteomics represents a potentially important tool for integrating protein-level information into the genome annotation process and improving genome annotation quality [17].

In this review, we highlight several ways in which high-throughput liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based proteomics can improve the quality of genome annotations, with emphasis on microbial genomes. Additionally, we highlight the value that could be gained by accompanying genome sequencing projects with even a modest set of high-throughput tandem mass spectrometry (MS/MS)-based proteomics experiments.

THE GENOME ANNOTATION PROCESS AND ITS CHALLENGES

The collective process of identifying genes (structural annotation) and assigning function to each gene (functional annotation) is commonly referred to as genome annotation. Protein-coding genes, which for the most part dictate biological function, comprise a small fraction of higher eukaryotic genomes, <25% of the fly and worm genome and an even smaller fraction of the human genome (<5%), making the identification of coding sequences against the ubiquitous background of non-coding sequences difficult. Further complicating the identification of protein-coding genes and their correct genomic structure is the high frequency of alternative splicing in most eukaryotic genes. Structural annotation in prokaryotes is far from being a trivial matter in spite of the compact nature of prokaryotic genomes and the usual absence of introns. While the identification of all possible open reading frames (ORFs) longer than a chosen threshold in a DNA sequence is a straightforward exercise, the decision as to which ORFs represent true coding genes that are expressed and code for proteins is not an insignificant endeavour with challenges in fundamental areas such as determining the precise start and stop site of a gene, accurately predicting short genes and determining a stop codon that represents an alternative amino acid rather than a true stop site, to name a few. Such challenges are only exacerbated and compounded when dealing with the annotation of the often larger and more complex genomes of eukaryotic organisms.

Likely protein-coding genes in genomic DNA sequences are identified using a range of computational tools referred to as automated genome annotation 'pipelines' that combine information from non-comparative (*ab initio*) and comparative (sequence similarity) methods. The TIGR CMR [18], GenDB [19] and BASys [20] represent commonly used pipelines in prokaryotic genome annotation. In a typical microbial genome annotation, raw DNA sequence is searched with *ab initio* microbial gene prediction programs such as GLIMMER [21, 22] or CRITICA [23] to predict protein-coding sequences. *Ab initio* gene prediction programs are designed to use statistical properties of ORFs such as G+C rich regions, codon usage and splice site consensus sequences to identify genes. In addition, the DNA sequence is also searched against the non-redundant database of publicly available proteins using the BLAST algorithm [24–26]. Integration of evidence from both methods leads to identification of the set of predicted protein-coding genes. For functional assignment predicted protein sequences are subjected to series of similarity searches and sequence analysis. These include searches against the COG database [27] to find putative orthologs in other completed genomes, against the TIGRFAM [28] and PFAM [29] databases for protein family analysis, against PROSITE [30] for sequence motif analysis, and against the protein localization prediction software PSORT [31]. Query sequence analysis with SignalIP [32] for signal peptide prediction, TMHMM [33] for the prediction of alpha helical trans-membrane regions and PSIPRED [34] for predicted secondary structure is also included.

The Ensembl [35], NCBI [36] and USCS [37] 'pipelines' are commonly used in eukaryotic genome annotation. In a typical mammalian genome annotation, known protein sequences from the genome of

interest or evolutionarily close relatives are aligned against the genome towards predicting protein-coding genes and their genomic structure. Additionally, for genomes with rich expressed sequence libraries, such as the human genome, cDNA derived for the genome of interest are aligned against the genome towards predicting protein-coding genes and their genomic structure. In the majority of sequenced genomes where expressed sequence libraries are not as well developed as those for the human genome, cDNA and protein alignments against the genome are often complemented by dual-genome *de novo* gene predictors such as SGP2 [38] and TWINSCAN [39]. These use results of alignments between two evolutionarily related genomes to modify the gene prediction scores produced by underlying single-genome *ab initio/de novo* gene prediction programs, under the assumption that regions conserved in the sequence will tend to correspond to protein-coding regions from homologous genes. Single-genome *ab initio* gene prediction programs such as GENSCAN [40] and GENEID [41] typically play important roles in annotation on occasions where no appropriate homologous genome exists and expressed sequence libraries offer minimal coverage of the expressed genome, by complementing cDNA and protein alignment methods.

The overwhelming majority of both prokaryotic and eukaryotic sequenced genomes lack the rich cDNA libraries associated with the human genome and are not as well curated. Thus, predictions of protein-coding regions for the majority of annotated genomes are heavily weighted on *de novo* gene prediction programs. Even in the human genome, with its deep expressed sequence libraries, while cDNA and protein alignment methods may identify protein-coding genes and provide verification for their transcription, and translation, respectively. These methods are likely to be biased against predicting genes expressed in a restricted manner or at very low levels, providing an incomplete picture of the coding sequences of the human genome. This is filled in using computational gene prediction tools whose results require experimental validation.

While *de novo* gene prediction programs have proven useful in eukaryotic genome annotation, in the human genome for example, they are estimated to predict the correct gene structure only 50% of the time [42]. This number is modestly higher in eukaryotes with compact genomes such as that of *A. thaliana*, where up to two-thirds of the time they are estimated to predict the correct gene structure [43]. In light of this, the need to verify the protein-coding gene predictions made by these computational tools in the eukaryotic annotation process becomes clear.

De novo gene prediction programs are able to predict protein-coding genes in a prokaryotic genome with a much higher accuracy compared to their performance on eukaryotic genomes, owing to the lack of introns and high gene density in prokaryotic genomes. However, challenges associated with determining the precise start and stop site of a gene, accurately predicting short genes and determining a stop codon that represents an alternative amino acid rather than a true stop site, to name a few, still remain. In a recent re-analysis of 143 annotated prokaryotic genomes Nielsen and Krogh [44] observed that in some genomes up to 60% of the genes may have been annotated with a wrong start codon, especially in the GC-rich genomes. They also observed that a significant fraction of the genomes analysed had been over-annotated due to a lack of discrimination between short random ORFs and real genes. This highlights the concern of propagating databases with inaccurate gene predictions; an issue that is set to worsen with the explosion in the number of prokaryotic sequencing efforts which will likely rely exclusively on *de novo* prediction programs for subsequent annotation. In light of these facts, the need to verify the protein-coding gene predictions made by these computational tools in the prokaryotic annotation process becomes apparent.

Experimentally validating predicted protein-coding genes ideally would entail isolating and sequencing a full length cDNA for the prediction, and then providing evidence that the cDNA is translated into a protein. Current methods for verifying the existence and genomic structure of predicted protein-coding genes involve the systematic RT-PCR and direct sequencing of gene predictions, as described by Wu and colleagues [45], which advanced techniques initially pioneered by Miyajima *et al.* [46], Das *et al.* [47] and Guigo *et al.* [48]. These expression-based validation techniques may be able to predict that a possible protein-coding gene is expressed or not; they importantly cannot provide the evidence that an expressed gene is translated into a protein. In addition, results of the RT-PCR can be biased by the initial genome annotation. For example, if the reading frame for a gene predicted from the

annotation process is incorrect, then the PCR primer designed is not going to target the correct and/or complete gene sequence.

Currently, the best option for independently and unambiguously identifying at least an important subset of the protein-coding genes in a genome is to perform a systematic analysis of the naturally expressed protein complement of the genome (the proteome) and then work backwards to the parent genomic sequence. Until the advent of high-throughput LC-MS/MS–based proteomics, this option remained technically impractical to using low-throughput and labour-intensive peptide sequence analysis techniques/procedures (e.g. N-terminal sequence analysis of peptides by Edman degradation or two-dimensional gel electrophoresis-based approaches).

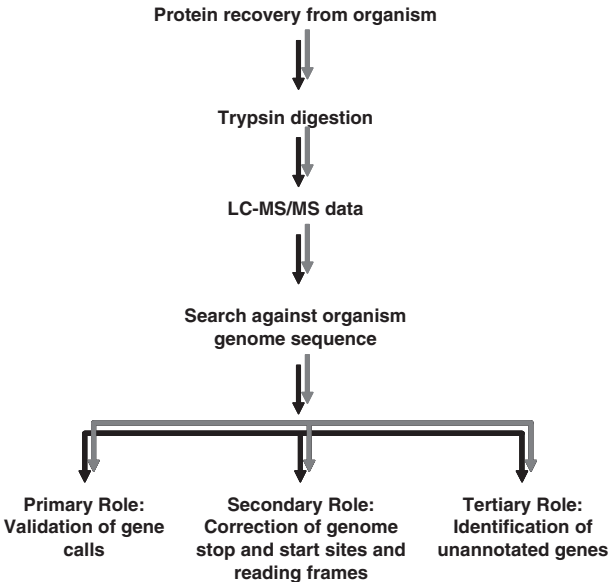

Figure 1: Workflow of mass spectrometry-based proteomics as applied to genome annotation. Following protein extraction, proteins are subjected to tryptic digestion, producing tryptic digest mixture. Tryptic digest mixture is analysed by capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS). MS/MS peptide spectra searched against specific organism genome sequence validating and correcting genomic annotations, as well as identifying novel protein-coding genes.

MS-BASED PROTEOMICS APPLIED TOWARDS GENOME ANNOTATION

In high-throughput LC-MS/MS-based proteomics, protein mixtures are digested with proteases, and the resulting peptides are typically first separated by multidimensional LC and then analysed by MS/MS [15, 16]. Each MS/MS spectrum is a measure of fragment masses, ideally from a single peptide sequence (~6–50 amino acids from an enzymatically digested protein). This set of mass values is analogous to a 'fingerprint' that identifies the peptide. Interpretation of the MS/MS peptide spectra is performed either (i) by using algorithms such as X! tandem [49], SEQUEST [50] or Mascot [51] to compare measured masses against a set of theoretical masses of possible protein sequences or (ii) less commonly, by *de novo* analysis, which does not depend on any prior knowledge of the possible sequences [15, 16]. Similar to searching MS/MS spectra against a set of predicted protein sequences, it is also possible to identify the protein-coding genes in a genome by searching MS/MS spectra against a six-frame translation of the genomic DNA sequence, thereby precluding the inherent biases derived from gene prediction methods. However, the exponential increase in peptide search space generated by the six-frame translation of large genomes results in increased search times. The elapsed search time in searching MS/MS spectra against a six–frame translation of the genomic DNA sequence is inversely proportional to both the speed and number of central processing units deployed, thus for efficient searching a networked cluster of processors is typically employed. A generalized strategy for mapping peptides identified by MS onto an existing annotated genome is outlined in Figure 1.

The idea of searching MS/MS spectra against nucleic acid sequences was first demonstrated by Yates and colleagues [52], in attempts to integrate data obtained from cDNA and genomic sequencing projects with biochemical studies. In this approach the nucleic acid sequence representing the 'genomic' database is translated in all six reading frames, and then queried with MS/MS spectra to identify protein-coding genes. Early proteomic analysis of the bacterium *H. influenzae* by Link and coworkers [53] represents the first report of a whole bacterial genomic sequence being queried with MS/MS spectra in this manner to identify protein-coding genes. The feasibility of querying MS/MS spectra against large eukaryotic genomic sequences in similar fashion was later demonstrated by Kuster [54] and Choudary [55] for the *A. thaliana* and human genomes, respectively. The results from these analyses produced a more confident annotation of corresponding genomes through the confirmation

of a set of predicted genes, identification of novel genes, independent validation of hypothetical ORFs and correction of erroneous gene predictions.

Current gold-standard expression-based techniques employed to verify the existence of predicted genes involve the systematic RT-PCR and direct sequencing of predicted protein-coding genes. While these methods can indicate that a predicted gene is expressed, they are not able to determine more importantly if the expressed gene is translated into a protein. Only proteomics can unambiguously determine if the expressed gene is translated into a protein. High-throughput LC-MS/MS-based proteomics directly measures protein fragments. The resulting peptide sequences confirm the existence of a subset of naturally occurring protein products from a specific genome (in a fashion that is not biased by genome annotation) and serve to validate an annotation, i.e. a protein-coding gene. Thus LC-MS/MS-based proteomics can facilitate genome annotation efforts if adopted regularly. In addition, the high-throughput nature of the LC-MS/MS-based proteomics techniques makes this technology cost-effective and readily applicable to the automated genome annotation process.

VALIDATION OF PREDICTED GENES AND DETECTION OF NOVEL GENES

There are now several studies in the literature where predicted genes have been validated at the protein level. For example, Jaffe and coworkers [56] used LC-MS/MS-based proteomics to validate a majority of the predicted genes in the genome of the bacterium *Mycoplasma pneumoniae*. In this work, peptides detected in a whole-cell lysate of *M. pneumoniae* were mapped onto its genome and these 'peptide hits' were extended to ORFs bound by traditional genetic signals to generate what they referred to as a 'proteogenomic map'. Using this proteogenomic annotation approach, the identity of many of the protein-coding genes of the *Mycoplasma mobile* genome [57] and the *Mycobacterium smegmatis* genome [58] were validated. In addition to validating predicted genes, this approach can be used to detect novel genes. Peptides that map to genomic regions outside the boundaries of previously annotated genes represent evidence of novel genes (or exons) or extensions of their predicted termini. LC-MS/MS-based proteomics revealed the existence of several new ORFs in the *M. tuberculosis* [59] and *M. pneumoniae* [56] genomes that were not originally predicted by genomic methods.

A number of studies have used LC-MS/MS-based proteomics to validate predicted genes at the translation level in plant (*A. thaliana*) [54], insect (*Anopheles gambiae*) [60] and human [55, 61–63] genomes. The complexities arising from the small fraction of coding sequence within higher eukaryotic genomes and the high frequency of alternative splicing in most eukaryotic genes lead to an increased rate of erroneous gene predictions, as such, there is an even greater need for experimental verification of the predicted protein-coding genes in eukaryotic genomes. The ability to detect novel genes in complex eukaryotic samples also has been demonstrated in much the same way. For example, LC-MS/MS-based proteomics has provided translational level evidence for several novel exons or genes and the extension of known exons in the *S. cerevisiae* [64], *D. melanogaster* [65], *A. gambiae* [60] and human [63] genomes.

PROTEOMIC VALIDATION OF HYPOTHETICAL OPEN READING FRAMES

Hypothetical genes usually represent a significant portion, i.e. ~30–50%, of all genes in a genome, and the rapidly growing number of hypothetical proteins with each newly sequenced genome is one of the emerging challenges of modern biology. Hypothetical and conserved hypothetical genes are predicted genes that either do not have any known homologs or are homologous to other hypothetical genes in other closely related organisms. Since the prediction of these genes is based entirely on the presence of a putative start, putative stop and upstream promoter region, the likelihood that they are erroneous predictions is higher and consequently the need for experimental validation is greater.

The current crop of gene prediction algorithms trained on proteobacteria datasets (as proteobacteria genomes were the first to be sequenced) has enabled relatively robust and accurate gene prediction in proteobacteria; however, as more distantly related organisms are sequenced, the level of accuracy for the current programs trained on proteobacteria datasets will markedly decrease leading to an increase in incorrect annotations of hypothetical genes. In light of these factors, it is imperative to experimentally

verify whether a hypothetical ORF is translated into a protein.

Through its ability to directly measure proteins LC-MS/MS-based proteomics can validate hypothetical genes at the protein level. In a global analysis of the *Deinococcus radiodurans* proteome, Lipton and coworkers [66] confirmed the expression of several genes previously annotated as hypothetical. The expression of several hypothetical and conserved hypothetical proteins has also been detected in a number of other prokaryotic genomes, including the *H. influenzae* [67], *Salmonellosis typhimurium* [14], *Salmonella typhi* [68], *Yersinia pestis* genome [69] and the *Shewanella oneidensis MR-1* genome [70, 71], which validates that these lower confidence gene predictions are accurate. Extension of this to higher order organisms with more complex eukaryotic genomes was demonstrated by Tanner and coworkers [63], who were able to confirm the translation of 224 hypothetical human proteins.

DETERMINATION OF PROTEIN START AND TERMINATION SITES

Another challenge for current gene prediction algorithms is determining the correct start position of a gene. In a recent re-analysis of 143 annotated prokaryotic genomes Nielsen and Krogh [44] observed that in some genomes up to 60% of the genes may have been annotated with a wrong start codon, especially in the GC-rich genomes. Accurate start site predictions better define intergenic spaces that may encode promoters and regulatory binding sites, which are critical elements in studies of transcriptional regulation. Cellular localization signals also are contained in start sites, which makes accurate start site predictions important for accurately determining the localization of proteins within a cell. Predicted translational start sites are typically confirmed by N-terminal sequencing using the Edman method; however, this is a low-throughput process, requiring isolation and is not amenable to the majority of proteins that have a 'blocked N-terminus'.

LC-MS/MS-based proteomics represents an approach for experimentally verifying predicted translational start sites, as in studies with *M. tuberculosis* [72], *S. oneidensis MR-1* [17] and *A. gambiae* [60] genomes. N-terminal peptide identifications are able to confirm predicted start sites and correct erroneously predicted start sites for multiple genes, as well as to discover the presence of rare start codons [17]. In a further advancement, Gevaert and coworkers [73, 74] recently described a method for employing a chemical modification strategy to enrich all the N-terminal peptides. As such, they were able to validate a large number of translational start sites in a single experiment.

Selenoproteins are proteins that incorporate the 21st amino acid selenocysteine. Selenocysteine is inserted into selenoproteins by a re-coding event known as codon re-assignment where the triplet codon Thymine-Guanine-Adenine (TGA), normally a stop codon, specifies selenocysteine insertion. TGA thus differs from all other codons in that it has a dual function, encoding selenocysteine and terminating translation. The alternative decoding of TGA is conferred by an mRNA stem-loop structure termed the selenocysteine insertion sequence (SECIS) element [75, 76]. In eukaryotes and archaea SECIS elements are located in 3′ untranslated regions [77, 78] and in bacteria are located immediately downstream of selenocysteine TGA codons [79, 80]. Although TGA has dual role as a stop codon or selenocysteine, available *de novo* gene prediction programs only interpret TGA as a stop codon leading to selenoprotein genes being misannotated or completely missed, and more broadly to the misannotation of the 3′ terminus of protein-coding genes. Recently, various computational approaches have been developed that have facilitated the prediction of selenoprotein genes in many eukaryotic genomes [81–87]. However, as yet no such comparable approaches have been developed for characterizing selenoproteins in prokaryotic genomes. LC-MS/MS by directly measuring proteins can be used to provide protein-level evidence of ORFs that contain an in-frame TGA codon representative of an alternative amino acid, thus highlighting an erroneous translation prediction.

APPLICATION TO COMPARATIVE BACTERIAL GENOMICS

The explosion in the number of available bacterial genome sequences, >450 and counting, has enabled the realization of the concept of comparative bacterial genomics. Comparative genomics is based on the idea that DNA sequences conserved between species are often those that encode functional and regulatory elements preserved from a common ancestor responsible for the essential biological processes of this common ancestor. Conversely,

inter-species differences in the comparative analysis may indicate DNA sequences that encode functional and regulatory elements driving functional adaptation [88, 89]. Consequently comparative bacterial genomics has been used to gain insights into the evolution of bacterial species and identification of potentially important novel genes [90]. However, given the fact that a gene will not always produce a gene product, i.e. protein, due to for example post-transcriptional regulatory mechanisms [91] and evolutionary silencing mechanisms [92] it is important that the expression of genes identified in comparative genomic studies be verified at the protein level. This however, is rarely done due to the low-throughput, labour-intensive and resource-heavy methods required to do so. High-throughput LC-MS/MS-based proteomic measurements provide a means to experimentally demonstrate the existence of genes identified in comparative genomic studies at the level of translation in a manner that alleviates most of the above limitations.

VERIFICATION OF EXISTENCE OF SPLICE VARIANTS AT THE PROTEIN LEVEL

Alternative splicing is an important molecular mechanism that allows a single gene to produce multiple protein isoforms, thereby playing a major role in the production of complex proteomes with a broad range of functional diversity. The resulting changes in amino acid sequence generated by alternative splicing allow a single gene to produce proteins that have different binding properties, cellular localization, stability and enzymatic activity [93]. The biological importance of this phenomenon is demonstrated in the regulation of programmed cell death, where a number of apoptotic genes are alternatively spliced generating protein isoforms often with either pro- or anti-apoptotic effects [94]. It is estimated that up to 70% of human genes undergo alternative splicing [93, 95, 96]. Given the importance and preponderance of alternative splicing events, alternative splicing detection cannot be neglected in the annotation process of a new genome. Current *ab initio* and sequence similarity methods used in structural gene prediction, however, face challenges in accurately predicting alternatively spliced gene structures. For example *ab initio* programs, which define predicted gene structures as an optimal prediction that is most probable according to its underlying probabilistic model, identify alternative splicing by searching for suboptimal predictions with probabilities very close to the optimal. This approach however is very debatable as alternatively spliced gene structures can be very different from the initial predicted gene structure. In addition there is need to distinguish between real splice variants and false positives, as many alternative predictions can always be made for any sequence, a specificity issue that can be resolved using orthogonal validation methods [97]. Alignment-based methods identify alternative splicing by aligning expressed sequence tags (ESTs) and cDNAs to the genomic sequence. This however is limited by the availability of transcribed sequences, non-uniformity of protocols and the labour-intensive and expensive nature of the procedures required [97–99]. Current high-throughput techniques for the validation of alternative splicing predictions by *ab initio*/bioinformatic methods involve the use of DNA microarray experiments to detect alternative splicing, followed by further validation using RT-PCR. These expression-based methods, however, only reveal that the splice variant it expressed, and provide no information on whether the expressed spliced variant is translated into a protein. By directly measuring protein fragments, with the resulting peptide sequences confirming the existence of a subset of naturally occurring protein products from a specific genome (in a fashion that is not biased by genome annotation), LC-MS/MS-based proteomics can verify the existence of splice variants at the protein level. For example, Tanner and colleagues [63] using LC-MS/MS were able to discover or confirm at the translation level over 40 alternative splicing events in the human genome.

HOW TO INTEGRATE PROTEOMICS WITH GENOME ANNOTATION

The initial annotation of the prokaryotic *Shewanella baltica* OS195 genome can be used to further illustrate the effectiveness of using proteomics data integrated with genomic annotation. Likely protein-coding DNA sequences in *S. baltica* OS185 were computationally predicted (identified) using either Glimmer or Critica. Overall protein-coding genes predicted by both algorithms are similar; however, a small, yet significant number of protein-coding genes are predicted by one algorithm and not the other, which demonstrates the potential ambiguity in the

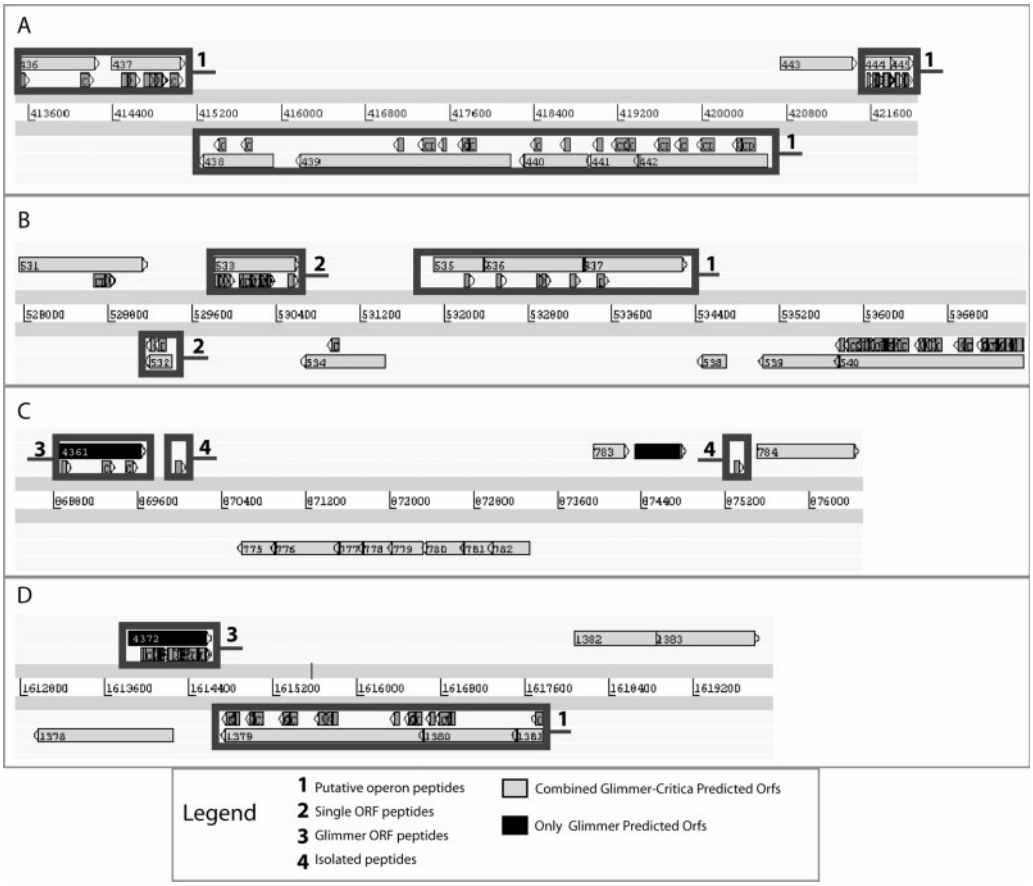

Figure 2: Illustrations depicting the use of LC-MS/MS data to validate computationally predicted genes viewed with the program artemis. Numbered light grey rectangles represent gene calls made by the program Critica and the numbered black rectangles represent gene calls made by the program Glimmer. The smaller blocks associated with the overlaying light grey or black numbered rectangles are peptide sequences obtained through tandem mass spectrometry, i.e. proteomics observed sequences. In cases where gene calls by Critica and Glimmer are identical, only the Critica representation is shown for clarity. Externally numbered dialogue boxes are used to highlight specific type of observations. Illustration of three likely operons including small ORFs with multiple peptides observed by LC-MS/MS. Illustrates both an operon and multiple single ORF proteins including small ORFs. An illustration of Glimmer's only calls an isolated peptide identifications. Isolated peptide identifications without associated gene predictions may be novel ORFs or falsely identified peptides. A Glimmer's only identification has multiple mass spectrometry-derived peptides (small blocks) mapped onto gene on forward strand predicted by Glimmer but none mapped to gene on reverse strand predicted by Critica. A high resolution color version of this figure is available on request (proteomics@pnl.gov).

annotation process. This situation highlights the problem of contradictory gene predictions that result from the use of different gene prediction algorithms providing even for a simple prokaryotic genome. The situation is magnified several fold for more complex eukaryotic genomes, where gene predictions are further complicated by the presence of introns and exons.

LC-MS/MS-based proteomics data provides an excellent opportunity to resolve these disparities by validating gene predictions and correcting over-predictions from the aggregated results of multiple gene prediction algorithms. Figure 2 illustrates the use of proteomic data for confirming/validating *S. baltica* OS185 genes predicted by both Critica and Glimmer. In cases where gene calls by Critica and Glimmer are identical, only the Critica representation is shown for clarity. The overlaying of experimentally observed peptides from the predicted coding regions shows good agreement in the various panels of Figure 2, this information could be used in a relatively automated fashion with the development of new tools.

Small ORFs are among the most difficult genomic features to predict and are often either missed in the annotation process due to conservative calls or over-represented due to over-calling. By providing direct evidence of expression, proteomics data confirms the existence of these small proteins without introducing extraneous coding regions as shown in Figure 2A and B. Figure 2C illustrates a scenario whereby proteomics resolved an apparent disparity between gene predictions made by Critica and Glimmer. The region between 868 800 and 869 600 is predicted by Glimmer to be a protein-coding DNA sequence (gene 4361); however, this region is missed by Critica. Three peptides identified by LC-MS/MS match this gene, thus confirming the prediction by Glimmer. Another example is shown in Figure 2D. Here, Critica predicts a protein-coding region between 1 612 800 and 1 614 400 (gene 1378). Glimmer also predicts a protein-coding region between 1 612 800 and 1 614 400 (gene 4372), but on the opposite strand. As depicted, the multiple small blocks that represent peptide sequences obtained through LC-MS/MS map onto the gene in the forward direction between 1 612 800 and 1 614 400, i.e. the gene predicted by Glimmer and represented by the black rectangle 4372. This observation confirms the gene prediction by Glimmer and indicates that the prediction made by Critica was likely an over-prediction, a chronic problem with most gene prediction algorithms.

Previous studies have shown that different sources of gene evidence can be combined to improve a final genome annotation [42]. Similarly, additional data obtained from LC-MS/MS analyses can be integrated with genome annotation pipelines to validate ambiguous gene calls, start sites and coding frames. For example, by incorporating MS/MS data as an additional line of evidence in the gene prediction program GENEID [41], Tanner and coworkers [63] added an additional 863 correctly identified human exons to their predictions.

METAGENOME ANNOTATION OF MICROBIAL COMMUNITIES

Microbial organisms play key roles in a variety of processes that range from balancing the composition of the atmosphere to fighting disease. In their natural environment, these organisms rarely function in isolation, but rather in the context of diverse microbial communities. Thus, an understanding of the structure and activities in microbial communities depends on the ability to sample genomic information from all member organisms. However, currently unculturable microbial organisms comprise the majority of organisms in most environments on earth. In light of this fact, culture-independent methods are essential to understand the genetic diversity, population structure and ecological roles of the majority of member organisms of microbial communities [100, 101].

Metagenomics has emerged as a powerful tool for culture-independent genomic analysis of a population of microbial organisms. In metagenomics, the genomic DNA is extracted directly from microbial communities in environmental samples, which negates the need to culture organisms under study and maintains potentially critical community and environmental interactions [101]. However, community sequencing efforts are hampered by the complexity introduced from a combination of many organisms in a single sample. This complexity creates challenges in assembling community genomes due to the increased size of the genome compared to single microbes and the redundancy of metabolic pathways within a community of organisms.

The value of LC-MS/MS-based proteomics to metagenomics was initially highlighted in a study by Ram and coworkers [102] that used LC-MS/MS-based proteomic methods to evaluate gene expression, identify key activities, and examine partitioning of metabolic functions in a natural acid mine drainage (AMD) microbial biofilm community; an environment with relatively limited complexity. A recent [103] LC-MS/MS-based proteomics study that helped reveal the importance of inter-population genetic exchange towards adapting to environmental conditions in an AMD microbial biofilm community, further demonstrated the potential of integrated community genomics and proteomics.

SUMMARY AND FUTURE DIRECTIONS/CHALLENGES

As the post-genomic era gains momentum, it has become increasingly clear that a DNA sequence alone is insufficient to provide an understanding of complex molecular and cellular biological processes. The fact that the bulk of current validation methods use information derived from an annotated genome and that not all ORFs are translated, i.e. one cannot independently and unambiguously

determine whether a predicted ORF is translated into a protein, represents an untenable situation for genome sequencing/annotation.

LC-MS/MS-based proteomics measures proteins directly, providing a high-throughput means of verifying at the level of translation the existence of a subset of naturally occurring protein products from a raw DNA sequence. In addition to validating predicted genes at the level of translation, MS/MS data also can be used to detect novel genes, confirm translation of hypothetical proteins, accurately determine translational start and stop sites, verify the existence of splice variants at the level of translation, correct erroneous annotations and provide information for incorporation into gene prediction algorithms to enhance gene prediction. In light of these tangible benefits and facts, we strongly recommend that every genome sequencing/annotation project be complemented with a standard proteomics effort.

In looking towards the future, the need for improved data mining and informatics tools and instrumentation is recognized as a key challenge that confronts the application of proteogenomic methods to genome annotation. As genome sequencing/annotation projects begin to include high-throughput LC-MS/MS datasets as a necessary complement to results from gene prediction programs, it is imperative that new functionalities be built into both current and new data mining and informatics tools that take full advantage of the value-added information provided by MS/MS datasets. For example, there is a need for visualization tools with the ability to display start sites only in the context of overlaid MS/MS data to accurately define and correct erroneously predicted translational start sites. Another example of an area of need is auto-detection of non-standard features, such as peptides that map over stop sites, which would enhance the efficiency of annotating unusual codon usage. With development, these tools will more intelligently use data that are available from proteomics experiments to fully annotate the questions that arise during annotation.

While the principle of searching tandem MS spectra data against six-frame translated genomes to experimentally validate predicted protein-coding genes has been demonstrated in both prokaryotes and eukaryotes, as previously described, at the present the technique faces technical challenges in its application to complex eukaryotic genomes. One challenge arises from the enormous database size of six-frame translated complex eukaryotic genomes which are orders of magnitude larger than the protein sequence databases traditionally employed in MS/MS peptide spectra analysis. For example, the human proteome has an estimated size of 25 Mb residues in comparison to six-frame translation of the human genome estimated to be 6 Gb residues [63]. Scaling up to larger databases in the context of the current search strategy employed in proteogenomic annotation methods makes searches extremely long to the point of being impractical. In addition, it also results in increased false-positive counts as the false discovery rate scales with the increasing database size, greatly decreasing sensitivity. Key to meeting these challenges are improved informatics and instrumentation that when combined, can improve search speed and the accuracy/sensitivity of the existing methods. The current method of choice to overcome the twin challenges of decreased search speeds and decreased sensitivity has been to reduce or restrict the peptide search space. Tanner and colleagues [63] created the exon graph database, a compact representation of all human putative exons, splice variants and polymorphisms constructed from the EST database and GENEID exon predictions, and searched this database instead of searching a six-frame translated genome directly. Significantly reducing their database size to 134 Mb residues and resulting in increased search speeds and sensitivity. Sevinsky and colleagues [104] developed the GENQUEST technique which uses isoelectric focusing and accurate mass to reduce the peptide search space in searching tandem MS spectra data against six-frame translated human genomes. This produced search times and sensitivity that appeared comparable to searching protein sequence databases with tandem MS spectra data.

While the above examples represent advances towards enabling efficient searching of tandem MS spectra data against raw genomic sequences from complex eukaryotes, further improvements are still needed to make this search routine less problematic and more widely applicable. In contrast, searching tandem MS spectra data against six-frame translated genomes of prokaryotes and certain simple eukaryotes remains a robust and efficient search routine that offers immediate benefits to improving genome annotation, and we advocate that whenever possible each genome sequencing project be complemented by a set of LC-MS/MS-based proteomics experiments.

Key Points

- State-of-the-art genome annotation combines *ab initio* predictions with sequence similarity data to predict protein-coding regions, the vast majority of these predictions are rarely experimentally validated.
- Most current techniques used to validate the existence of predicted protein-coding genes are based on expressed RNAs; however, the presence of an RNA sequence cannot independently and unequivocally determine whether a predicted gene is translated into a protein.
- LC-MS/MS-based proteomics directly measures peptides arising from expressed proteins, which allows for the direct verification of observed coding regions of a genomic sequence and leads to improved genome annotations.
- Additional insights for genome annotation from LC-MS/MS-based proteomics include indentifying hard to predict protein-coding regions and verifying the existence of splice variants.

Acknowledgements

The authors gratefully acknowledge the contributions of Dr Margie Romine and Penny Colton for their assistance in preparing this publication. This research was supported by the Genomics:GtL Program, Office of Biological and Environmental Research, US Department of Energy (DOE) at PNNL under grant (ER63232-1018220-0007203) and the National Institute of Allergy and Infectious Diseases NIH/DHHS through interagency agreement Y1-AI-4894-01. Samples were analysed in the Environmental Molecular Sciences Laboratory (a DOE national scientific user facility) located at Pacific Northwest National Laboratory in Richland, Washington.

References

1. Fleischmann RD, Adams MD, White O, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 1995;**269**:496–512.
2. Fraser CM, Gocayne JD, White O, *et al*. The minimal gene complement of Mycoplasma genitalium. *Science* 1995;**270**: 397–403.
3. Cole ST, Brosch R, Parkhill J, *et al*. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 1998;**393**:537–44.
4. Goffeau A, Barrell BG, Bussey H, *et al*. Life with 6000 genes. *Science* 1996;**274**:546,563–547.
5. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 1998;**282**:2012–8.
6. Myers EW, Sutton GG, Delcher AL, *et al*. A whole-genome assembly of Drosophila. *Science* 2000;**287**:2196–204.
7. The *Arabidopsis* Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000;**408**:796–815.
8. Waterston RH, Lindblad-Toh K, Birney E, *et al*. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**:520–62.
9. Gibbs RA, Weinstock GM, Metzker ML, *et al*. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004;**428**:493–521.
10. Sodergren E, Weinstock GM, Davidson EH, *et al*. The genome of the sea urchin Strongylocentrotus purpuratus. *Science* 2006;**314**:941–52.
11. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
12. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51.
13. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;**431**:931–45.
14. Adkins JN, Mottaz HM, Norbeck AD, *et al*. Analysis of the Salmonella typhimurium proteome through environmental response toward infectious conditions. *Mol Cell Proteomics* 2006;**5**:1450–61.
15. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
16. Washburn MP, Wolters D, Yates JR, 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;**19**:242–7.
17. Gupta N, Tanner S, Jaitly N, *et al*. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Res* 2007;**17**:1362–77.
18. Peterson JD, Umayam LA, Dickinson T, *et al*. The comprehensive microbial resource. *Nucleic Acids Res* 2001; **29**:123–5.
19. Meyer F, Goesmann A, McHardy AC, *et al*. GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res* 2003;**31**:2187–95.
20. Van Domselaar GH, Stothard P, Shrivastava S, *et al*. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res* 2005;**33**:W455–9.
21. Delcher AL, Harmon D, Kasif S, *et al*. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;**27**:4636–41.
22. Salzberg SL, Delcher AL, Kasif S, *et al*. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 1998;**26**:544–8.
23. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 1999; **16**:512–24.
24. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
25. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
26. Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet* 1993;**3**:266–72.
27. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**:631–7.
28. Haft DH, Loftus BJ, Richardson DL, *et al*. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* 2001;**29**:41–3.
29. Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;**32**:D138–41.
30. Hulo N, Sigrist CJ, Le Saux V, *et al*. Recent improvements to the PROSITE database. *Nucleic Acids Res* 2004;**32**:D134–7.
31. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 1999;**24**:34–6.

32. Nielsen H, Engelbrecht J, Brunak S, *et al*. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 1997;**10**:1–6.
33. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 1998;**6**: 175–82.
34. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;**16**:404–5.
35. Hubbard T, Andrews D, Caccamo M, *et al*. Ensembl 2005. *Nucleic Acids Res* 2005;**33**:D447–53.
36. Maglott D, Ostell J, Pruitt KD, *et al*. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005;**33**: D54–8.
37. Karolchik D, Baertsch R, Diekhans M, *et al*. The UCSC genome browser database. *Nucleic Acids Res* 2003;**31**:51–4.
38. Parra G, Agarwal P, Abril JF, *et al*. Comparative gene prediction in human and mouse. *Genome Res* 2003;**13**:108–17.
39. Korf I, Flicek P, Duan D, *et al*. Integrating genomic homology into gene structure prediction. *Bioinformatics* 2001;**17**(Suppl. 1):S140–8.
40. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;**268**:78–94.
41. Parra G, Blanco E, Guigo R. GeneID in Drosophila. *Genome Res* 2000;**10**:511–15.
42. Guigo R, Flicek P, Abril JF, *et al*. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* 2006;**7**(Suppl 1):S21–31.
43. Allen JE, Pertea M, Salzberg SL. Computational gene prediction using multiple sources of evidence. *Genome Res* 2004;**14**:142–8.
44. Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* 2005;**21**:4322–9.
45. Wu JQ, Shteynberg D, Arumugam M, *et al*. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. *Genome Res* 2004;**14**:665–71.
46. Miyajima N, Burge CB, Saito T. Computational and experimental analysis identifies many novel human genes. *Biochem Biophys Res Commun* 2000;**272**:801–7.
47. Das M, Burge CB, Park E, *et al*. Assessment of the total number of human transcription units. *Genomics* 2001;**77**: 71–8.
48. Guigo R, Dermitzakis ET, Agarwal P, *et al*. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc Natl Acad Sci USA* 2003;**100**:1140–5.
49. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004;**20**:1466–7.
50. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;**5**:976–89.
51. Perkins DN, Pappin DJ, Creasy DM, *et al*. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999;**20**: 3551–67.
52. Yates JR, 3rd, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* 1995;**67**:3202–10.
53. Link AJ, Hays LG, Carmack EB, *et al*. Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143. *Electrophoresis* 1997;**18**:1314–34.
54. Kuster B, Mortensen P, Andersen JS, *et al*. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 2001;**1**:641–50.
55. Choudhary JS, Blackstock WP, Creasy DM, *et al*. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* 2001;**1**:651–67.
56. Jaffe JD, Berg HC, Church GM. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 2004;**4**:59–77.
57. Jaffe JD, Stange-Thomann N, Smith C, *et al*. The complete genome and proteome of Mycoplasma mobile. *Genome Res* 2004;**14**:1447–61.
58. Wang R, Prince JT, Marcotte EM. Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* 2005;**15**:1118–26.
59. Jungblut PR, Muller EC, Mattow J, *et al*. Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics. *Infect Immun* 2001;**69**: 5905–7.
60. Kalume DE, Peri S, Reddy R, *et al*. Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics* 2005;**6**:128.
61. Desiere F, Deutsch EW, Nesvizhskii AI, *et al*. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol* 2005;**6**:R9.
62. Fermin D, Allen BB, Blackwell TW, *et al*. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol* 2006;**7**:R35.
63. Tanner S, Shen Z, Ng J, *et al*. Improving gene annotation using peptide mass spectrometry. *Genome Res* 2007;**17**: 231–9.
64. Oshiro G, Wodicka LM, Washburn MP, *et al*. Parallel identification of new genes in Saccharomyces cerevisiae. *Genome Res* 2002;**12**:1210–20.
65. Brunner E, Ahrens CH, Mohanty S, *et al*. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotechnol* 2007;**25**:576–83.
66. Lipton MS, Pasa-Tolic L, Anderson GA, *et al*. Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc Natl Acad Sci USA* 2002;**99**: 11049–54.
67. Kolker E, Makarova KS, Shabalina S, *et al*. Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae. *Nucleic Acids Res* 2004;**32**: 2353–61.
68. Ansong C, Yoon H, Norbeck AD, *et al*. Proteomics analysis of the causative agent of typhoid fever. *J Proteome Res* 2008; **7**:546–57.
69. Hixson KK, Adkins JN, Baker SE, *et al*. Biomarker candidate identification in Yersinia pestis using organism-wide semiquantitative proteomics. *J Proteome Res* 2006;**5**: 3008–17.
70. Elias DA, Monroe ME, Marshall MJ, *et al*. Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. *Proteomics* 2005;**5**:3120–30.

71. Kolker E, Picone AF, Galperin MY, *et al*. Global profiling of Shewanella oneidensis MR-1: expression of hypothetical genes and improved functional annotations. *Proc Natl Acad Sci USA* 2005;**102**:2099–104.
72. Rison SC, Mattow J, Jungblut PR, *et al*. Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of *Mycobacterium tuberculosis*. *Microbiology* 2007; **153**:521–8.
73. Aivaliotis M, Gevaert K, Falb M, *et al*. Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J Proteome Res* 2007;**6**:2195–204.
74. Gevaert K, Goethals M, Martens L, *et al*. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol* 2003;**21**:566–9.
75. Low SC, Berry MJ. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends Biochem Sci* 1996;**21**:203–8.
76. Thanbichler M, Bock A. The function of SECIS RNA in translational control of gene expression in *Escherichia coli*. *EMBO J* 2002;**21**:6925–34.
77. Berry MJ, Banu L, Chen YY, *et al*. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3′ untranslated region. *Nature* 1991;**353**: 273–6.
78. Wilting R, Schorling S, Persson BC, *et al*. Selenoprotein synthesis in archaea: identification of an mRNA element of *Methanococcus jannaschii* probably directing selenocysteine insertion. *J Mol Biol* 1997;**266**:637–41.
79. Liu Z, Reches M, Groisman I, *et al*. The nature of the minimal 'selenocysteine insertion sequence' (SECIS) in *Escherichia coli*. *Nucleic Acids Res* 1998;**26**:896–902.
80. Zinoni F, Heider J, Bock A. Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. *Proc Natl Acad Sci USA* 1990;**87**: 4660–4.
81. Castellano S, Morozova N, Morey M, *et al*. *In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep* 2001;**2**:697–702.
82. Castellano S, Novoselov SV, Kryukov GV, *et al*. Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep* 2004;**5**:71–7.
83. Kryukov GV, Castellano S, Novoselov SV, *et al*. Characterization of mammalian selenoproteomes. *Science* 2003; **300**:1439–43.
84. Kryukov GV, Kryukov VM, Gladyshev VN. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem* 1999;**274**:33888–97.
85. Lambert A, Lescure A, Gautheret D. A survey of metazoan selenocysteine insertion sequences. *Biochimie* 2002;**84**:953–9.
86. Lescure A, Gautheret D, Carbon P, *et al*. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J Biol Chem* 1999;**274**: 38147–54.
87. Martin-Romero FJ, Kryukov GV, Lobanov AV, *et al*. Selenium metabolism in Drosophila: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J Biol Chem* 2001;**276**:29798–804.
88. Hardison RC. Comparative genomics. *PLoS Biol* 2003;**1**: E58.
89. Tirosh I, Bilu Y, Barkai N. Comparative biology: beyond sequence analysis. *Curr Opin Biotechnol* 2007;**18**:371–7.
90. Callister SJ, McCue LA, Turse JE, *et al*. Comparative bacterial proteomics: analysis of the core genome concept. *PLoS ONE* 2008;**3**:e1542.
91. Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* 2005;**21**:399–404.
92. Mira A, Pushker R. The silencing of pseudogenes. *Mol Biol Evol* 2005;**22**:2135–8.
93. Stamm S, Ben-Ari S, Rafalska I, *et al*. Function of alternative splicing. *Gene* 2005;**344**:1–20.
94. Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. *Mol Cell* 2005;**19**:1–13.
95. Brett D, Pospisil H, Valcarcel J, *et al*. Alternative splicing and genome complexity. *Nat Genet* 2002;**30**:29–30.
96. Kriventseva EV, Koch I, Apweiler R, *et al*. Increase of functional diversity by alternative splicing. *Trends Genet* 2003;**19**:124–8.
97. Foissac S, Schiex T. Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* 2005;**6**:25.
98. Johnson JM, Castle J, Garrett-Engele P, *et al*. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 2003;**302**:2141–4.
99. Modrek B, Lee C. A genomic view of alternative splicing. *Nat Genet* 2002;**30**:13–19.
100. Allen EE, Banfield JF. Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 2005;**3**:489–98.
101. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 2004; **68**:669–85.
102. Ram RJ, Verberkmoes NC, Thelen MP, *et al*. Community proteomics of a natural microbial biofilm. *Science* 2005;**308**: 1915–20.
103. Lo I, Denef VJ, Verberkmoes NC, *et al*. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 2007;**446**:537–41.
104. Sevinsky JR, Cargile BJ, Bunger MK, *et al*. Whole genome searching with shotgun proteomic data: applications for genome annotation. *J Proteome Res* 2008;**7**:80–8.