Journal of Bacteriology and Parasitology

ISSN: 2155-9597

Armengaud, et al., J Bacteriol Parasitol 2011, S3
DOI: 10.4172/2155-9597.S3-001

Research Article Open Access

Microbial Proteogenomics, Gaining Ground with the Avalanche of Genome Sequences

Jean Armengaud[1]*, Céline Bland[1], Joseph Christie-Oleza[2] and Guylaine Miotello[1]

[1]*CEA, DSV, IBEB, Lab Biochim System Perturb, Bagnols-sur-Cèze, F-30207, France*
[2]*School of Live Science, University of Warwick, Coventry CV4 7AL, United Kingdom*

Abstract

A consensus for defining the fundamental unit of biological diversity, the species, has not yet been reached for prokaryotes. Although high-throughput molecular tools are now available to assess the microbial diversity, estimating the total number of species of bacteria and archaea on Earth is still a challenge due to the huge amount of low-abundant species present in environmental samples. Ever since the first whole cellular genome sequenced, the one from Haemophilus influenzae in 1995, more than seven thousand complete genomes have been reported. The avalanche of genome sequences is resulting in an exceptional documentation of representatives of numerous taxa. While annotation of these genomes has gained in accuracy with new gene prediction tools, proteogenomics has proved to help in discovering new genes, identifying the true translational initiation codon of coding domain sequences and characterizing maturation events at the protein level such as signal peptide processing. Beside this structural annotation, proteogenomics can also give rise to significant insights into the function of proteins. Basically, proteogenomics consists in obtaining massive protein sequence data by means of large shotgun proteomic strategies and the use of high-throughput tandem mass spectrometry. Such experimental data is then used for improving genome annotation. Unexpected results such as the reversal of gene sequences in different bacteria or the use of non-canonical start codons for translation in Deinococcus species are only some of the numerous corrections documented so far. Today, the proteogenomic analysis of a given set of representatives that fully covered the tree of life would result in a better ground for accurate annotation of novel strains. This would improve comparative genomics studies and could be of help for assessing in what way closely-related species are differing.

Microbial Genome Sequencing and Annotation Opened a Comprehensive View on How a Living Cell Functions

The first whole genome of a living cellular organism ever sequenced and annotated was that of *Haemophilus influenzae* Rd KW20 [1]. This gamma-proteobacterium was chosen for such technical challenge because of its relatively small genome (1.83 Mbp), its low GC ratio (38.2 %) and being a model of interest as pathogen. This bacterium was shown to be a secondary pathogen in influenza and can be synergistic with the influenza virus. It is one of the leading causes of meningitis in young children. It may also cause septicemia, chronic bronchitis and diverse inflammations such as otitis (inflammation of the middle ear) or sinusitis (inflammation of the sinus cavity).

While genome sequencing and annotation of such bacteria is now done in a few days, (even for bigger and more complex genomes) the arduous task done in the 90's was remarkable. Such improvements have been possible because of important innovations in sequencing techniques and the strategies used. Firstly, the use of fluorescent dyes for labeling nucleotides when preparing samples for the Sanger sequencing reaction instead of the former radioactive compounds was an important step in order to reduce the number of samples to be handled as well as the needed security precautions. The use of capillary electrophoresis instead of large polyacrylamide gels was a second round of improvement. Regarding the strategies, genomists quickly shifted from a chromosome walking methodology requiring design and synthesis of numerous oligonucleotides and consumption of a significant amount of time to a shotgun strategy where randomly-generated and overlapping fragments are cloned into a plasmid and sequenced with universal oligonucleotides. Moreover, after a period where scientists invested important efforts to get a precise chromosome sequence and manual annotation of each coding domain sequence (CDS), a new era is now open in which most of the genome information is obtained without losing time with a fastidious finishing steps [2].

Whole Genome Shotgun (WGS) sequencing projects are incomplete genomes or incomplete chromosomes that are being sequenced by a whole genome shotgun strategy. The genome information of a WGS project is currently presented in pieces or contigs (overlapping reads). Novel sequencing techniques have been proposed over the 00's turning genome sequencing into a real high-throughput approach. The so-called "next-generation sequencing" based on 454 FLX (Roche), SOLiD (Applied Biosystems) or Solexa (Illumina) technologies has triggered an impressive avalanche of genomic data. With such a high-throughput potential, a novel level of complexity (metagenomics) could be analyzed with the sequencing of total DNA extracted from environmental samples without any cultivation or isolation of a given set of bacteria [3]. The large diversity of living microbes is then probed in samples from different origins.

Microbial genome sequencing and annotation opened a comprehensive view on how a living cell functions. A total of 1,650 bacterial, 117 archaeal, and 37 eukaryotic annotated genomes have been now released (2011/10/09) and to date, more than five thousand genomes have been sequenced and their annotations are currently

***Corresponding author:** Jean Armengaud, Laboratoire de Biochimie des Systèmes Perturbés, CEA Marcoule, DSV, iBEB, SBTN, LBSP, F-30207 BAGNOLS-SUR-CEZE, France, Tel: +33(0)466796802; Fax: +33(0)466791905; E-mail: jean.armengaud@cea.fr

Received October 24, 2011; **Accepted** November 15, 2011; **Published** November 18, 2011

Citation: Armengaud J, Bland C, Christie-Oleza J, Miotello G (2011) Microbial Proteogenomics, Gaining Ground with the Avalanche of Genome Sequences. J Bacteriol Parasitol S3-001. doi:10.4172/2155-9597.S3-001

under progress. It is important to note that genome sequences are now confidently established while sequences of the first sequenced organisms were not. Because of this, re-sequencing and re-annotation of previous models have been proposed [4].

Regarding the annotation tools, several softwares have been refined taking into account CDS predictors and gene conservation [5-8]. The most recent tools allow a rapid annotation and comparison of genomes from multiple isolates of the same species, i.e. pan-genomes concept [9]. Pipelines integrating these different annotation tools are now available allowing a first genome annotation draft within only some hours.

Obtaining the genome information of model microorganisms led to considerable development of new scientific fields over the last fifteen years: genomics, comparative genomics, structural genomics, transcriptomics and proteomics are some examples. Proteomics consists in the analysis of proteins at the whole-genome scale. Proteomics was first initiated by chemical Edman-sequencing of the N-termini of a limited number of proteins. Based on whole-genome sequencing, identification of a protein was possible after establishing a few residues of either the N-terminus or an internal sequence of a given protein. Nowadays, proteomics relies on data obtained by mass spectrometry which can be easily interpreted if the protein sequences encoded on the genome are known [10]. *De novo* sequencing of a given protein is tedious and a time-consuming task when starting the interpretation of mass spectrometry data from scratch by using only theoretical considerations. In the early 90's, protein identification was performed on the basis of peptide mass fingerprint recorded by MALDI-TOF [11]. This consisted in resolving the proteome by 2D-gel electrophoresis to get the protein of interest "purified" as a single homogeneous spot, proteolyzing this protein with trypsin and recording the exact masses of the resulting peptides. The comparison of the mass pattern with those predicted for all the proteins contained in a database was rather simple and straightforward. Combination of only four masses determined with an error tolerance below 100 ppm resulted in most cases to protein identification.

However, in the late 90's, a new mass spectrometry approach was developed to directly establish the sequence of each of the peptides of a complex mixture avoiding a 2D-gel electrophoresis step. In this case, digested peptides obtained from a complex protein mix are first resolved by reverse chromatography on a reverse phase, ionized by electrospray, analyzed by a first mass spectrometry analyzer, then softly fragmented into smaller peptides by collision with neutral gas molecules and the masses of these resulting entities are recorded with a second analyzer (or even the same one). This strategy is called tandem mass spectrometry and gives rise to MS/MS spectra representative of a given peptide sequence [12].

As shown in Figure 1, hundreds of proteins can now be processed in a single shotgun analysis [13]. The resulting MS/MS experimental data are compared to the theoretical spectra that may be obtained for all the theoretical peptides comprised within the database, taking into account theoretical or statistical fragmentation rules deduced from fragmentation of model peptides [14]. It is usually admitted that detection of two peptide sequences within such an approach allows certifying the presence of the corresponding protein. The ratio of false-positive identification can be evaluated by querying the same dataset against decoy databases. Today tandem mass spectrometers are so accurate and rapid that a sample comprising a thousand of proteins can be comprehensively analyzed in a few hours and semi-quantified if the dynamic range is not too large [15,16].

Why is Genome Annotation not Yet Fully Comprehensive and Accurate?

The annotation of a genome consists in identifying its coding sequences transcribed into functional RNAs or translated into proteins. Also, the annotation can include important genomic signals for transcription, translation and associated regulatory mechanisms. Based on their characterization by molecular biology techniques over the last 40 years, most of these signals can be now searched with automatic procedures but a reliable and comprehensive view cannot be yet achieved because of the complexity of these mechanisms. Moreover, genome annotation should give a precise description of the function of each of the proteins encoded by the chromosome. Cataloguing as much information as possible on these proteins would be the idealistic annotation objective. These two levels of annotation are respectively named structural annotation (description of the location of each key item) and functional annotation (description of the function) and are intimately linked by nature. Softwares designed for this multi-level annotation are now developed together with novel database structures for the integration of heterogeneous biological information [17].

For several decades, tremendous efforts of the biologist community have been invested in order to decipher the function of numerous proteins. Indeed, the most important cellular mechanisms are now known and well described in terms of their functional players for several model micro-organisms. However, despite these efforts, a quite large ratio of proteins are until now either poorly characterized, remain with an unknown function, or purely hypothetical. Among the latter group we find orphan proteins identified by genome inspection and which present no detectable similarities with any other protein within the existing databases. Their existence could be considered as doubtful because of the Life history. It is now well established that a new protein has a higher probability to emerge from the duplication of a previous coding sequence and derive by point mutations, deletions or insertions of amino acid residues, rather than to appear from scratch [18,19]. When the derived protein acquires a new function that starts to impact the fitness of the organism it can then be maintained. Since a protein should fold in the proper way to give a stable three-dimensional structure for a correct function, obtaining a new function from scratch is statistically highly improbable but cannot be a priori rejected.

Because we are still failing to identify these specific proteins that could be responsible for the specific traits in a given microbial species [20], it seems difficult today to establish by a sole genome inspection the exact number of proteins encoded in a given genome. Some model microorganisms have been subjected to substantial efforts in terms of sequencing. As an example, *Escherichia coli*, the workhorse of most molecular biology laboratories and an important medical species, is outstandingly well documented with more than 49 annotated genomes released and 540 additional genome sequences available (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi; 2011/10/09). Such redundant genome information can be then used to better annotate conserved proteins by just comparing closely related genomes.

As genomes are now annotated by automatic systems based on comparative genomics, novel nucleic acid sequences that do not share any similarities with already known sequences are subjected to numerous errors [21]. For example, the two first *Deinococcus species* ever sequenced, *Deinococcus radiotolerans* BAA-816 [22] and *Deinococcus geothermalis* DSM1130 [23], revealed numerous annotation errors when a third Deinococcus genome sequence (from *Deinococcus deserti*) was established [24]. It is demonstrated that

annotation errors are generally maintained and propagated once new relatively-close genomes are being annotated. Databases are currently saturated of non-informative sequences and complete genomes. Indeed, their use needs specific precautions.

Possible reasons for annotation errors were recently discussed [25]. Besides these reasons, an increase of sequencing errors in newly sequenced genomes could be introduced due to the most recent technologies based on ultra-short read data sets that show some specific trends of wrong base calls [26]. Second-generation annotation systems, which combine multiple gene-calling programs with similarity-based methods, perform in a better way than the first annotation tools. However, novel improvements of these annotation systems are urgently needed. Our recent work related to the systematic identification of translation start codons in *Deinococcus deserti* [27] revealed that a high ratio (approximately 20%) of proteins were erroneously annotated in terms of N-terminus. An in silico genome analysis estimates that the prediction of start codons is frequently erroneous (10-20%), reaching up to 60 % in some GC-rich prokaryotic genomes [28].

Proteogenomics, the Use of Proteomic Data to Improve Genome Annotation

Current methodologies for microbial proteome in-depth analysis rely on the use of high-resolution tandem mass spectrometers coupled to high-pressure liquid chromatography systems for resolving peptide complex mixtures. Such experimental set-up is currently able to record more than 18,000 MS/MS high-quality spectra in one hour, and thus usually detect more than 5,000 distinct peptides. Consequently, it is possible to first resolve complex mixture of proteins upon either their molecular weights by 1D SDS-PAGE or their isoelectric points by OFFGel electrophoresis. Then, these proteins can be proteolyzed with trypsin before the resulting peptide mixtures are characterized by nanoLC-MS/MS (Figure 1). A comprehensive list of the most abundant proteins and their quantitation by such approach give crystal-clear evidences of their synthesis.

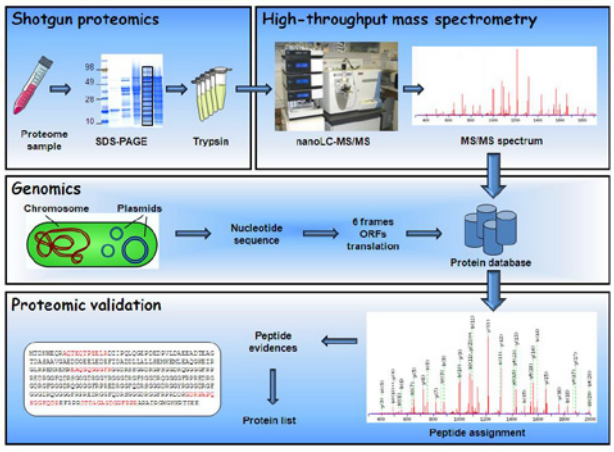

Figure 1: Shotgun proteomics for a proteogenomic approach: most common strategy and current tools.
The most classical approach in shotgun proteomics consists in i) fractionation of proteins upon their molecular weights by 1D SDS-PAGE, prior to ii) trypsin proteolysis of the protein fractions (polyacrylamide bands), and iii) peptide identification by nanoLC-MS/MS with high throughput hybrid mass spectrometers. MS/MS data are then assigned to peptide sequences by specific searches against publicly available databases or homemade databases comprising all the possible protein sequences. For proteogenomics, the database is made of a six-frame translation of the nucleic acid sequence. Such approach typically results in assignment of tens of thousands MS/MS spectra and identification of hundreds of proteins with two or more non-redundant peptides.

Proteogenomics consists in the detection of specific proteotypic peptides to reveal the existence of genes encoding proteins annotated as hypothetical or simply missed during genome annotation. For this, the MS/MS spectra are assigned to peptide sequences using a database made of a six-frame translation of the whole genome, and so containing mostly unlikely protein sequences. The truly existing proteins are thus extracted from this large list once several mass spectrometry evidences are recorded. Their sequences can be then mapped onto the loci of the genome that encode them, resulting in an informative "proteogenomic map" as reviewed [29-31].

In addition to prove the real existence of an hypothetical gene, proteogenomics allows the detection of unannotated genes, reversal of reading frames, establishment of correct translational start sites (protein N-terminus), detection of programmed frameshifts as well as characterization of post-translational modifications such as signal peptide maturation, amino acid lateral chain change and presence of inteins. The large studies of post-translational modifications in *Shewanella oneidensis* MR-1 [32] or *Salmonella typhimurium* 14028 [33] illustrates the benefits of such approaches.

Deinococcus deserti VCD115 is another example where proteogenomics has been massively used [24,27]. This bacterium was isolated from a mixture of sand samples collected in the Sahara Desert in Morroco and Tunisia [34]. To better understand the adaptation of this microorganism to harsh conditions encountered in hot arid deserts, its complete genome sequencing and annotation were carried out. Its genome consists of a 2.8 Mb chromosome and three large megaplasmids, totaling 3.8 Mbp. A large shotgun proteomic analysis was carried out at the primary stage while the genome was still being sequenced. A set of 11,129 unique peptides were recorded by tandem mass spectrometry that led to the confident identification of 1,348 proteins. A large number of non-predicted genes by the two annotation softwares used were revealed by the proteogenomic analysis. More surprisingly, the reversal orientation was observed for eleven incorrectly predicted genes [24]. Figure 2 shows such proteogenomic map established for a specific locus of *Deinococcus deserti* chromosome. Interestingly, the same locus comprised two important annotation errors: a small open reading frame (ORF) was unannotated despite the use of several annotation softwares and a predicted ORF was annotated when the opposite strand was the real polypeptide coding region (Figure 2). Both annotations could be corrected because of the detection of several peptides by mass spectrometry. As shown in Figure 2, Deide_19965 is a short ORF of 95 amino acids that do not present any detectable similarities to known protein sequences. Deide_19972 is an ORF conserved amongst Deinococcus species. However, the wrong annotation of this specific locus was also observed for the other species and should be corrected [24]. More recently, the identification of N-termini of *D. deserti* proteins on a very large scale was also carried out [27]. For this, the proteome was labeled with a succinimide reagent, namely N-Tris (2,4,6-trimethoxyphenyl) phosphonium acetyl succinimide (TMPP), that selectively derivatizes protein N-termini. After proteolysis with trypsin or chymotrypsin proteases, the resulting peptides were analyzed by nanoLC-MS/MS with a LTQ-Orbitrap XL high resolution mass spectrometer. A set of 664 N-terminal peptidic sequences were listed, leading to the correction of 63 translation

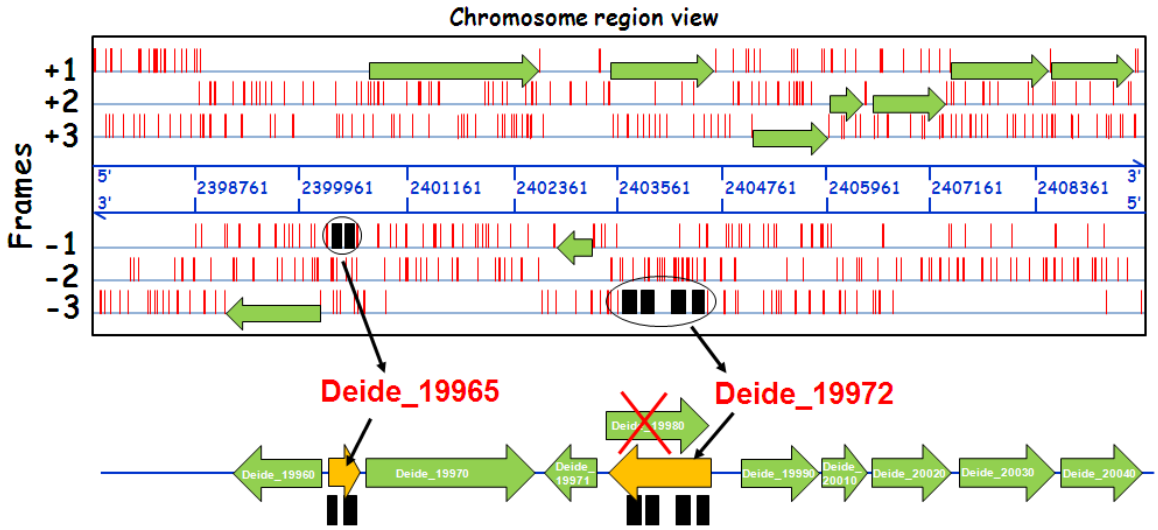

Figure 2: Mass spectrometry-based identification of genome miss-annotations exemplified with a specific *Deinococcus deserti* locus.
The alignment of mass spectrometry identified peptides directly onto the nucleic acid sequence gives a proteogenomic map. This map shows the real location of the coding domain sequences in the genome and highlights different cases of miss-annotation that should be further validated with an in-depth sequence analysis. Here, the different peptides assigned to MS/MS spectra recorded by tandem mass spectrometry are indicated with black rectangles directly onto their corresponding reading frame (forward or reverse strand). The figure shows a specific chromosome locus from *Deinococcus deserti* [24]. Genome position (NC_012526) is indicated in blue. Translational STOP codons are indicated with vertical red bars. Coding sequences which have been previously annotated automatically are indicated by green arrows. Peptides detected by tandem mass spectrometry are pointing at the presence of two novel miss-annotated proteins Deide_19965 (a short polypeptide of 95 residues coded on the -1 frame) and Deide_19972 (a polypeptide of 311 residues coded on the -3 frame that differs from the wrongly annotated Deide_19980 protein on the +1 frame).

initiation codons in the genome of *D. deserti*. Usually, protein translation is initiated from the ATG initiation codon, and in a less extent GTG or TTG codons may be used. Noteworthy, experimental evidences in the proteome of *D. deserti* indicated that some mRNA translations are initiated from ATC or CTG non-canonical codons [27]. Additional studies may illustrate how annotation errors can be overcome by a proteomic-driven approach [33,35-38].

Moreover, it is possible to include more information into the proteogenomic annotation such as gene expression data. Recently, RNA deep-sequencing data [39-43] or more traditional experimental confirmation of gene transcription for specific loci [50] have been successfully used to better annotate genomes. As discussed previously [44], when allied to genomics and other omics techniques, proteomics can help in providing high quality genome annotations at a relatively low cost. The discovery and annotation of numerous small proteins using genomics, proteomics and transcriptomic data for the *Populus deltoides* tree exemplifies how strategies first developed for microorganisms are now currently used in higher eukaryotes [43].

Over the last years a trend has emerged with the combination of genomic and proteomic data at the primary annotation stage of a genome annotation project [30]. Proteomic MS/MS data can be recorded while the nucleic acid sequence is still being established. Jaffe and co-workers [45] were pioneers in reporting such innovative idea presenting the complete sequence of the small-genome of Mycoplasma mobile (0.78 Mbp) together with its proteome. We contributed to generalize this initiative with two large projects on the *Deinococcus deserti* bacterium [24] and the *Thermococcus gammatolerans* archaeon [38], with genomes of 3.85 Mbp and 2.05 Mbp respectively.

The Expected Tsunami of Genome Sequences and The Very Beginnings of Ortho-Proteogenomics and Comparative Proteogenomics

The first "next-generation" DNA sequencer, 454 FLX from the Roche company, was launched in 2005 and rapidly impacted the genomics landscape [46]. Only six years later we will embrace the third generation of sequencing technology that should generate sequences in higher orders of magnitude. The current avalanche of genome data should turn out quickly into a massive tsunami where thousands of new microbial genomes could be established every week. The structural annotation of these genomes will surely gain in accuracy and comprehensiveness because of novel homologous sequences and conserved pattern searches. The input of evolutionary constraints will certainly be stronger when various genomes of each bacterial genus will be sequenced.

Taking a parallel approach to comparative genomics, a methodology based on the comparison of genome sequences to gain insights into the function of proteins and genome evolution trends, comparative proteogenomics has been proposed in order to use the evolutionary constraints to fully interpret proteogenomic data and therefore, improve genome annotations [47]. Here, researchers take advantage of mass spectrometry data obtained from multiple genomes belonging to the same genus giving more confidence to the reannotations. As an example, while analyzing the proteome of a single microorganism, some proteins or some peptides signing a specific maturation event will not be identified with enough confidence to be taken into consideration for the structural reannotation of the corresponding genes. Also, the confidence of the so-called "one-hit wonders" (proteins identified by a unique peptide) is at a low level and these identifications are usually discarded because ambiguous. If such peptides and respective homologues are seen in the proteomes of several closely-related microorganisms, even with different sequences and therefore m/z ratio, then the confidence for this specific event increases. In other words, the same protein seen with only one peptide but in two different species has more chances of being valid than two false-positive events.

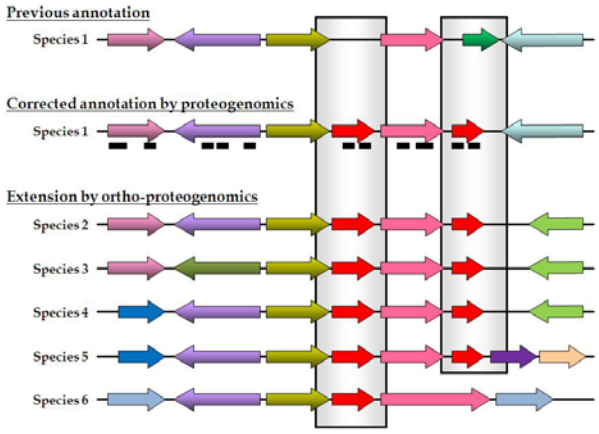

Figure 3: Ortho-proteogenomics: extending proteogenomic annotation to closely-related genomes.
Colored arrows represent coding ORFs in sequenced genomes from six related-species. Black rectangles beneath these ORFs in species 1 represent the peptides detected by tandem mass spectrometry through its proteome analysis. Two cases of miss-annotations were corrected in Species 1 and are indicated by red arrows. These corrections can be then extended to the other five species because of the conservation of the whole locus in most cases (grey vertical rectangles).

As a result, more confident exhaustive data will greatly improve the proteogenomic annotation of the different microorganisms under consideration, as experienced with the case-study of three *Schewanella* species [47].

In addition, proteogenomic data can be better exploited when combined with comparative genomics helping to refine annotation of multiple genomes. Certified annotation of one organism by proteogenomics experimental data can be applied to all orthogolous genes present in genomes from phylogenetically-related species. This is possible because of the evolutionary constraints that maintain specific gene structural signals if they are important along Life history (Figure 3). Gallien et al. [48] were the firsts to report such an extension of proteogenomic evidences obtained on one microorganism to the whole genus [48]. They tentatively called such extension "ortho-proteogenomics". They systematically determine the N-termini of numerous proteins from *Mycobacterium smegmatis*. The MS/MS data set collected on this proteome allowed to correct up to 19 % of the characterized start codons apart from identifying 29 missed annotated proteins. These corrections were then extended by sequence comparison to a set of 16 sequenced *Mycobacterium* species. A total of 4,328 re-annotations were obtained despite all the annotation efforts carried out during the numerous *Mycobacterium* genus sequencing projects. The same strategy has been applied to a set of 22 *Yersinia* genomes after an in-depth characterization of the *Yersinia pestis* KIM proteome [49].

Recently, we also extended this concept after analyzing the proteome of a marine *Roseobacter* species, *Ruegeria pomeroyi* DSS-3 [50]. The proteogenomic evidences obtained for *R. pomeroyi* DSS-3 were used to correct many annotation errors in a large number of species from distinct genera belonging to the *Roseobacter* clade. With this generalization concept in hands, we proposed that analyzing a given set of representative proteogenomes that fully covered the tree of life should result in a set of perfectly annotated genomes. This set of reference genomes could then establish the guidelines for the training of universal prediction tools for future genome annotation.

Conclusions

Novel sequencing technologies allow an extremely fast production of large sequence data sets. Their current evolution will generate in the near future an unprecedented avalanche of genome data. However, many annotation errors are still frequent and propagated throughout newly annotated genomes. Efforts to improve annotation pipelines are urgently required with the input of artificial intelligence for deciphering all possible transcription, translation, maturation and regulation consensus signals. Proteogenomics, the use of direct experimental evidences obtained at the protein level together with their maturation events, has proved reliable for a better structural genome annotation. Proteogenomics has been proposed to further help in discovering new genes, identifying the true translational initiation codon of each gene and characterizing protein maturation events. This already established technique, which intimately combines genomics and proteomics, also gives interesting insights into the function of many proteins as proteomics can systematically quantify proteins obtained from different physiological conditions [50]. Today, the analysis of a given set of representative proteogenomes that fully covered the tree of life would definitely improve existing and future annotated genomes. This would be possible if several state-of-the-art proteomic platforms contribute to generate these experimental data and if novel softwares are developed to methodically trace the proteomic evidences back to the genomic annotation level [51]. Quick achievements could be obtained in the framework of an international collaborative research project. Such research projects would result in a better ground for accurate comparative genomics and could be of help for assessing in what way closely-related species differ.

Acknowledgements

C Bland is supported by the Commissariat à l'Energie Atomique et aux Energies Alternatives and the Région Languedoc-Roussillon. JA Christie-Oleza was supported by a fellowship from the Fundación Ramón Areces. We thank the Commissariat à l'Energie Atomique et aux Energies Alternatives, the Agence Nationale de la Recherche (ANR-07-BLAN-0106-02), and the Région Languedoc-Roussillon (label « Chercheur d'Avenir Confirmé » 2010) for financial supports.

References

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd Science 269: 496-512.
2. Stothard P, Wishart DS (2006) Automated bacterial genome analysis and annotation. Curr Opin Microbiol 9: 505-510.
3. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5: e16.
4. Barbe V, Cruveiller S, Kunst F, Lenoble P, Meurice G, et al. (2009) From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. Microbiology 155: 1758-1775.
5. Kumar K, Desai V, Cheng L, Khitrov M, Grover D, et al. (2011) AGeS: a software system for microbial genome sequence annotation. PLoS One 6: e17469.
6. Stewart AC, Osborne B, Read TD (2009) DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics 25: 962-963.
7. Vallenet D, Engelen S, Mornico D, Cruveiller S, Fleury L, et al. (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. Database (Oxford) 2009: bap021.
8. Yu C, Zavaljevski N, Desai V, Johnson S, Stevens FJ, et al. (2008) The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. BMC Bioinformatics 9: 52.
9. Angiuoli SV, Hotopp JC, Salzberg SL, Tettelin H (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinformatics 12: 272.

10. Patterson SD, Aebersold RH (2003) Proteomics: the first decade and beyond. Nat Genet 33: 311-323.

11. Pusch W, Kostrzewa M (2005) Application of MALDI-TOF mass spectrometry in screening and diagnostic research. Curr Pharm Des 11: 2577-2591.

12. Nesvizhskii AI, Vitek O, Aebersold R (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4: 787-797.

13. Han X, Aslanian A, Yates JR 3rd (2008) Mass spectrometry for proteomics. Curr Opin Chem Biol 12: 483-490.

14. Sadygov RG, Cociorva D, Yates JR 3rd (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods 1: 195-202.

15. Clair G, Roussi S, Armengaud J, Duport C (2010) Expanding the known repertoire of virulence factors produced by *Bacillus cereus* through early secretome profiling in three redox conditions. Mol Cell Proteomics 9: 1486-1498.

16. Dedieu A, Gaillard JC, Pourcher T, Darrouzet E, Armengaud J (2011) Revisiting iodination sites in thyroglobulin with an organ-oriented shotgun strategy. J Biol Chem 286: 259-269.

17. Medigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158: 724-736.

18. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20: 1313-1326.

19. Tautz D, Domazet-Loso T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12: 692-702.

20. Cohan FM, Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. Curr Biol 17: R373-R386.

21. Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17: 429-431.

22. White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286: 1571-1577.

23. Makarova KS, Omelchenko MV, Gaidamakova EK, Matrosova VY, Vasilenko A, et al. (2007) *Deinococcus geothermalis*: the pool of extreme radiation resistance genes shrinks. PLoS One 2: e955.

24. de Groot A, Dulermo R, Ortet P, Blanchard L, Guerin P, et al. (2009) Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. PLoS Genet 5: e1000434.

25. Poptsova MS, Gogarten JP (2010) Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology 156: 1909-1917.

26. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.

27. Baudet M, Ortet P, Gaillard JC, Fernandez B, Guerin P, et al. (2010) Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. Mol Cell Proteomics 9: 415-426.

28. Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 21: 4322-4329.

29. Ansong C, Purvine SO, Adkins JN, Lipton MS, Smith RD (2008) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct Genomic Proteomic 7: 50-62.

30. Armengaud J (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol 12: 292-300.

31. Krug K, Nahnsen S, Macek B (2011) Mass spectrometry at the interface of proteomics and genomics. Mol Biosyst 7: 284-291.

32. Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, et al. (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res 17: 1362-1377.

33. Ansong C, Tolic N, Purvine SO, Porwollik S, Jones M, et al. (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella typhimurium*. BMC Genomics 12: 433.

34. de Groot A, Chapon V, Servant P, Christen R, Saux MF, et al. (2005) *Deinococcus deserti* sp. nov., a gamma-radiation-tolerant bacterium isolated from the Sahara Desert. Int J Syst Evol Microbiol 55: 2441-2446.

35. Rodriguez-Ortega MJ, Luque I, Tarradas C, Barcena JA (2008) Overcoming function annotation errors in the Gram-positive pathogen *Streptococcus suis* by a proteomics-driven approach. BMC Genomics 9: 588.

36. Batista JS, Torres AR, Hungria M (2010) Towards a two-dimensional proteomic reference map of *Bradyrhizobium japonicum* CPAC 15: spotlighting "hypothetical proteins". Proteomics 10: 3176-3189.

37. Zhao L, Liu L, Leng W, Wei C, Jin Q (2011) A Proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. BMC Genomics 12: 528.

38. Zivanovic Y, Armengaud J, Lagorce A, Leplat C, Guerin P, et al. (2009) Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea. Genome Biol 10: R70.

39. Frank S, Klockgether J, Hagendorf P, Geffers R, Schock U, et al. (2011) *Pseudomonas putida* KT2440 genome update by cDNA sequencing and microarray transcriptomics. Environ Microbiol 13: 1309-1326.

40. Larsen PE, Trivedi G, Sreedasyam A, Lu V, Podila GK, et al. (2010) Using deep RNA sequencing for the structural annotation of the Laccaria bicolor mycorrhizal transcriptome. PLoS One 5: e9780.

41. Legendre M, Santini S, Rico A, Abergel C, Claverie JM (2011) Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. Virol J 8: 99.

42. Tisserant E, Da Silva C, Kohler A, Morin E, Wincker P, et al. (2011) Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. New Phytol 189: 883-891.

43. Yang X, Tschaplinski TJ, Hurst GB, Jawdy S, Abraham PE, et al. (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res 21: 634-641.

44. Armengaud J (2010) Proteogenomics and systems biology: quest for the ultimate missing parts. Expert Rev Proteomics 7: 65-77.

45. Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, et al. (2004) The complete genome and proteome of *Mycoplasma mobile*. Genome Res 14: 1447-1461.

46. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24: 133-141.

47. Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, et al. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. Genome Res 18: 1133-1142.

48. Gallien S, Perrodou E, Carapito C, Deshayes C, Reyrat JM, et al. (2009) Ortho-proteogenomics: Multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res 19: 128-135.

49. Payne SH, Huang ST, Pieper R (2010) A proteogenomic update to *Yersinia*: enhancing genome annotation. BMC Genomics 11: 460.

50. Christie-Oleza JA, Fernandez B, Nogales B, Bosch R, Armengaud J (2011) Proteomic insights into the lifestyle of an environmentally relevant marine bacterium. ISME J [Epub ahead of print].

51. Castellana N, Bafna V (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics 73: 2124-2135.

This article was originally published in a special issue, **Microbial Genomics** handled by Editor(s). Dr. Luiz Carlos de Mattos, FARMERP, Brazil; Dr. Jesús Valdés Flores, CINVESTAV, USA; Dr. Shesheer Kumar, RAS Lifesciences, India