Computational approaches for the prediction of protein function in the mitochondrion

Toni Gabaldón

Am J Physiol Cell Physiol 291:C1121-C1128, 2006. First published 26 July 2006;
doi:10.1152/ajpcell.00225.2006

Am J Physiol Cell Physiol 291: C1121–C1128, 2006.
First published July 26, 2006; doi:10.1152/ajpcell.00225.2006.

Invited Review

Computational approaches for the prediction of protein function in the mitochondrion

Toni Gabaldón

Bioinformatics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain

Submitted 1 May 2006; accepted in final form 24 July 2006

Gabaldón, Toni. Computational approaches for the prediction of protein function in the mitochondrion. *Am J Physiol Cell Physiol* 291: C1121–C1128, 2006. First published July 26, 2006; doi:10.1152/ajpcell.00225.2006.—Understanding a complex biological system, such as the mitochondrion, requires the identification of the complete repertoire of proteins targeted to the organelle, the characterization of these, and finally, the elucidation of the functional and physical interactions that occur within the mitochondrion. In the last decade, significant developments have contributed to increase our understanding of the mitochondrion, and among these, computational research has played a significant role. Not only general bioinformatics tools have been applied in the context of the mitochondrion, but also some computational techniques have been specifically developed to address problems that arose from within the mitochondrial research field. In this review the contribution of bioinformatics to mitochondrial biology is addressed through a survey of current computational methods that can be applied to predict which proteins will be localized to the mitochondrion and to unravel their functional interactions.

genomic context; proteome

THE DETAIL AND ACCURACY with which we can model the system-level properties of the mitochondrion relies substantially on how comprehensive our knowledge is about its individual components and their corresponding interactions. In recent years, great progress has been achieved toward the identification of the complete set of proteins that perform their functions inside the mitochondrion, the so-called mitochondrial proteome. For instance, advances in subcellular proteomics have allowed the isolation of 615 proteins from human heart mitochondria (60), and the combination of experimental research with genomics and sequence analyses has expanded the set of mitochondrial proteins to nearly a thousand (13). Nevertheless, despite these efforts, a substantial fraction of the human mitochondrial proteome, estimated to contain about 1500 proteins (37), remains still unidentified. Moreover, a large fraction of the identified proteins have unknown functions, and for many of the rest the knowledge of their biological roles is just general.

During the last decade, the concourse of computational biology in mitochondrial research has been essential. For instance, bioinformatics tools have been widely used for predicting which proteins are targeted to the mitochondrion and for identifying their functional homologs. More recently, a number of novel computational techniques that integrate different sources of data and unravel new functional interactions among proteins have been developed (24). These techniques, known as context-based function prediction methods, are increasingly being used in the context of the mitochondrial proteome and have proven especially useful for the identification of novel disease genes. Here I survey the most prominent computational biology methods that are used in the field of mitochondrial biology, from the first sequence analysis algorithms that identify mitochondrial proteins, to the most sophisticated comparative genomics techniques that integrate different sources to predict functional interactions.

IDENTIFYING THE COMPLETE REPERTOIRE OF MITOCHONDRIAL PROTEINS

Perhaps one of the first computational approaches to be specifically applied to the mitochondrion was the development of sequence-based algorithms to detect mitochondrial targeting signals (12, 44). These methods are usually based on the detection of common properties of NH_2-terminal signal peptides present in many mitochondrial proteins. These common sequence features result from physical constrains imposed by the processes of recognition and translocation across the mitochondrial membranes. For instance, as a consequence of the net positive charge needed during the $\Delta\Psi$-driven import across the inner membrane (67), mitochondrial targeting peptides are enriched in positively charged residues and lack negatively charged ones. Additionally, they can form amphiphilic α-helices, which are used to bind the receptors at the mitochondrial outer membrane (1). Such physical parameters are computed by MitoProt (12) to derive a linear function, which is then compared with a cutoff for mitochondrial/non-mitochondrial localization prediction. Other methods, such as TargetP (19),

Address for reprint requests and other correspondence: T. Gabaldón, Bioinformatics Dept., Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler 16, 46013 Valencia, Spain (e-mail: tgabaldon@cipf.es).

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

SignalP (44), or Predotar (55), use neural network predictors that are trained on sets of proteins with known localization. Although these methods display a decent specificity, their use on a genomic scale can give rise to rather high false-positive rates, because the prior probability of a protein being mitochondrial is low. For instance, TargetP, with 91% specificity and 60% sensitivity, gives rise to a 69% false-positive rate, when applied at a genomic scale, since only 7% of the human proteome is localized to the mitochondrion (9). Moreover, a common limitation of all these methods is that they cannot detect proteins whose mitochondrial localization is mediated by internal signals (30). To avoid this, other computational methods have been developed that are independent of the presence of targeting signals. For instance, MITOPRED (31) is based on the Pfam domain occurrence patterns and the amino acid compositional differences encountered between mitochondrial and nonmitochondrial proteins. Yet another strategy to predict mitochondrial localization is based on the analysis of the phylogenetic profile of a protein, that is its pattern of presence/absence in a set of genomes (39). The idea behind this approach is that the endosymbiotic origin of mitochondria from within the alpha-proteobacteria would be reflected in the phylogenetic profiles of their proteomes, and, therefore, eukaryotic proteins with homologs in alpha-proteobacterial species are expected to have mitochondrial localization. Despite the great expectations raised by this original method, current knowledge on the evolution of the mitochondrial proteome makes it advisable to use this with caution, since most of the mitochondrial proteins do not originate from the alpha-proteobacteria, and a considerable fraction of proteins derived from this bacterial group have a nonmitochondrial localization (25, 26). Nevertheless, the presence of homologs in *Rickettsia prowazekii* is still used, in combination with other lines of evidence, as indicative for mitochondrial localization (9, 51).

For many years computational prediction was the only feasible technique to obtain a broad view of the mitochondrial protein complement, since experimental approaches to characterize protein localization were not amenable for use at a large scale. More recently, advances in experimental techniques, such as large-scale green fluorescent protein (GFP) tagging (32) and subcellular proteomics (70), are paving the way to the complete experimental characterization of the mitochondrial proteome. So far, quite comprehensive proteomics sets exist for human (60) and some model organisms such as mouse (40) and yeast (54). An obvious advantage of experimental techniques over computational predictions is that the former can be specifically applied to different tissues (21) or under different experimental conditions, thus obtaining a snapshot of the proteins localized to the mitochondrion in a given context. Moreover, proteomics are more informative in the sense that they provide quantitative measures of the abundance of each identified protein. Common pitfalls of proteomics techniques, however, are that they are usually biased toward abundant proteins and that they often miss proteins that are difficult to extract and analyze, e.g., integral membrane proteins.

An optimal solution to overcome the different limitations of the various methods is the integration of their results. Such an approach is used by the MitoP2 mitochondrial proteome database (51). This server integrates data from computational prediction and subcellular proteomics techniques, but also

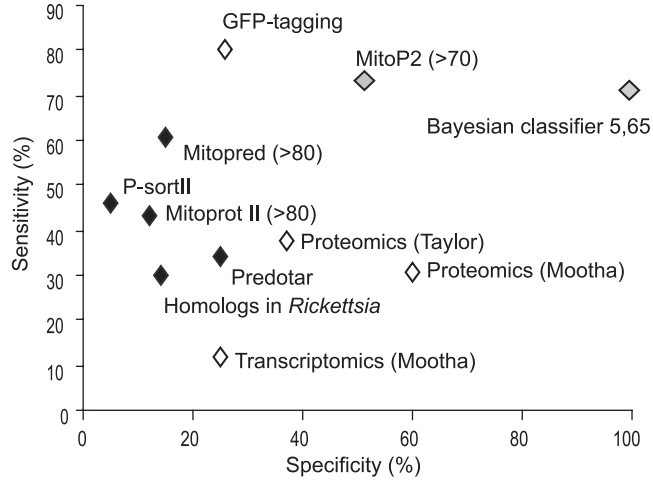

Fig. 1. Comparative analysis of the different methods used for predicting the human mitochondrial proteome. Sensitivity (true positives/ true positives + false negatives) and specificity (true negatives/ true negatives + false positives) are represented for each of the following methods: computational predictions (black diamonds), such as P-sort II (43), Mitoprot II (11), Predotar (55), Mitopred (31), and the presence of homologs in *Rickettsia* (51); experimental methods (white diamonds), such as green fluorescent protein (GFP)-tagging (48), transcriptomics of mouse mitochondria (40), and proteomics of human (60) and mouse (40) mitochondria; and finally, combined approaches (gray diamonds), such as MitoP2 (51) and the Bayesian classifier of Calvo et al. (9). In all cases the same reference sets of mitochondrial and nonmitochondrial proteins were used to compute specificity and sensitivity measures (taken from the supplementary material of Ref. 51).

results from other experiments that might be indicative of mitochondrial localization. These include, among others, mRNA expression profiles (40), phenotypes (6) and large-scale GFP-tagging screenings (48). All these different lines of evidence are subsequently combined in a single score, the MitoP2 score (52). This score considers the different specificities, that is, the percentage of proteins detected by each method (and their combinations) that is part of a reference data set. A recent implementation of MitoP2 (51), although still only applied to yeast mitochondria, uses support vector machines (SVMs) to combine all different data sets. SVMs are learning machines that can be trained to solve classification tasks (69). In the case of MitoP2 the SVM is trained, with a reference set, to classify proteins as mitochondrial or nonmitochondrial, according to a 20-dimensional input-vector that comprises the results for this protein in 20 different data sets. A similar integrative approach to predict human mitochondrial proteins was used by Calvo and co-workers (9). In this case they combined eight different sources of information and used a naïve Bayes classifier, called *Maestro*, trained on a reference data set of known mitochondrial proteins. Assuming independence between the different lines of evidence, this classifier employs Bayesian statistics to compute, for each protein, a likelihood of being mitochondrial. Figure 1 shows differences in sensitivity and specificity for most of the methods discussed, in all cases the same reference data set was used for the benchmark. As it can be seen, integrative approaches, such as that of MitoP2 or the Bayesian classifier are clearly superior to any other method used in isolation. The repertoire of mitochondrial proteins is thus rapidly increasing and several dedicated repositories provide listings of mitochondrial proteins, including sequence and

functional information (4, 13, 51). Considering the significant advances that have been achieved in the last few years, the full identification of the human mitochondrial proteome within the coming years seems a feasible goal.

HOMOLOGY-BASED FUNCTIONAL INFERENCE

Identifying the pieces that form the mitochondrial proteome constitutes only the first step toward the characterization of the system-level properties of the mitochondrion. To properly assemble the different mitochondrial components, it is necessary to first identify their functions. This is ideally done through experimental research, in which by subsequent assays one can functionally characterize a given protein or pathway. This process has been greatly accelerated by recent developments in high-throughput techniques that are able to produce experimental data for thousands of genes in one go. However, despite these efforts the experimental characterization of proteins is still a time consuming and expensive task. Fortunately, there are a number of computational techniques that can be exploited to assign function to experimentally uncharacterized proteins.

The classic and most widely used strategy to computationally annotate a protein consists of a transfer of knowledge from an experimentally annotated protein to its uncharacterized homologs, a technique called homology-based function prediction. A complete survey of homology-based function prediction methods is beyond the scope of this section and has been covered in specific reviews (53). Here, I will just provide a very brief overview, focusing on some cautionary remarks regarding the interpretation of the results.

The main advantage of homology-based function prediction resides in that it reduces the process of experimentally characterizing a protein to the much simpler task of finding an annotated homolog in a sequence database. For this purpose, there are a plethora of tools and websites that can be used. The most popular algorithms to detect significant similarities among protein sequences include Smith-Waterman (57) and BLAST (2). More sensitivity in the searches can be achieved by profile-based approaches such as Psi-Blast (3) or HMMER (17). The proliferation of user-friendly servers, in which nonspecialists can perform homology searchers, has facilitated the popularization of homology-based function prediction. It must be noted, however, that its use is not always straightforward and some caution must be taken when extrapolating functional annotations among proteins. This is particularly true when the comparisons involve sequences from distantly related species. Firstly, homologous proteins tend to share a common function at the molecular level, but this function can be performed in the context of completely different biological processes. For instance, two homologous protein kinases may trigger different signaling pathways and thus play a completely different biological role. Secondly, small changes in the amino acid sequence of a protein might result in significant variations of its function, e.g., a change in the substrate specificity of an enzyme. These, and other sources of errors might lead to incorrect annotations that can be rapidly propagated in sequence databases. It is, therefore, important to always consider the original source of the annotation provided by annotation frameworks such us Gene Ontology (28). Another useful advice is to consider not only the best hit in a sequence similarity search but to globally inspect a wider range of homologs down in the hit list. This will provide us with information on whether the function is conserved among that protein family. Another important consideration is to evaluate whether the sequence similarity extends over the full lengths of both sequences or if it is, otherwise, restricted to a given domain. The finding of partial-length homology can be useful if the function of the region of homology is known in one of the proteins. In this context, the search against domain databases such as Pfam (5) or SMART (36) can be of great help.

Finally, the accuracy of homology-based function prediction can be increased if orthology, rather than just homology, relationships are used. Two proteins are orthologous to each other when they evolved by speciation from a common ancestral sequence, in contrast to paralogs, which evolved by gene duplication (20, 22). Orthology is a relevant concept for function prediction since orthologs are, relative to paralogs, more likely to perform the same function. Orthology relationships should ideally be assessed through the detection of speciation and duplication events in a phylogenetic analysis (22). However, since phylogenetic reconstruction is a computationally heavy task, alternative methods, which only rely on sequence similarity searches, are more generally used. These include "best bi-directional hits" (33) and its multiple-genome extensions such as COG (59) or inparanoid (45).

The impact that homology-based function prediction has had in the annotation of proteins is beyond any doubt. However, it does not represent the definitive solution to the problem of the full functional characterization of mitochondrial proteins. After applying homology-based function prediction techniques, a great fraction of the mitochondrial proteome remains unannotated, and for many of the rest our knowledge on their function is just general. Fortunately, the genome era has inspired the development of novel computational methods that provide functional information that is complementary to that of homology-based function prediction (24). These so-called context-based methods are described in the next section.

CONTEXT-BASED FUNCTION PREDICTION

Context-based methods exploit genome comparisons to define the so-called genomic context of a gene. Here the term "context" is used to refer to any kind of information at a genomic scale, from the positions of the genes along the chromosomes, to their specific evolutionary history. If two genes share a particular context and this is significantly conserved during evolution, then a functional interaction between these two genes can be inferred. This functional interaction is usually of a different nature from that predicted by homology-based function prediction. While homology-based methods provide information on the function of a protein at a molecular level, e.g., its enzymatic activity, genomic-context methods provide information about the biological process in which that protein is playing a role, e.g., biochemical pathway. Thus ideally both approaches should be used in combination to increase the specificity of the functional predictions. Three main types of genomic context associations are generally used (Fig. 2): *1*) physical proximity in the genome, *2*) coevolution in terms of sequence and/or phylogenetic distribution, and *3*)

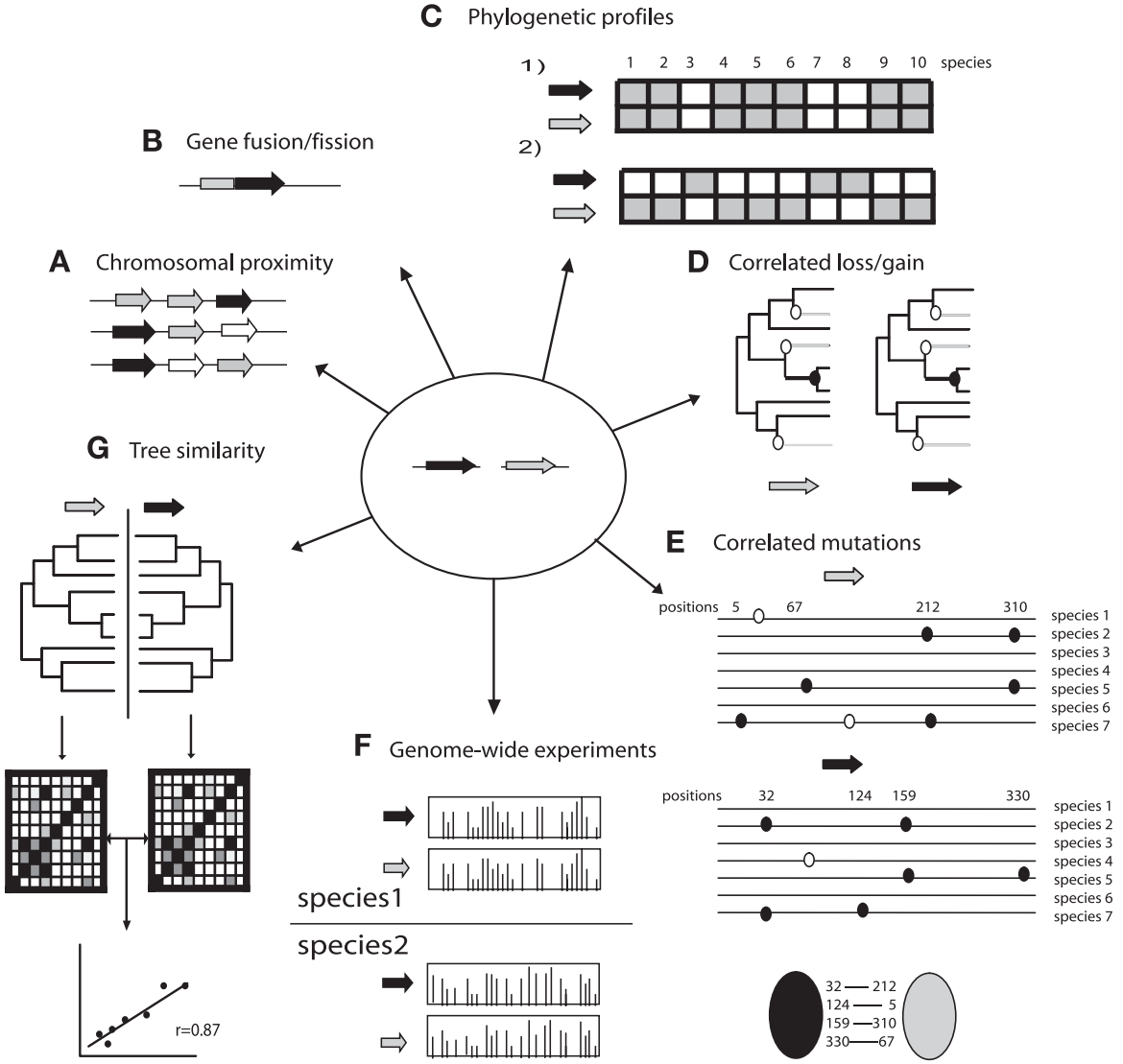

Fig. 2. Overview of the most widely used context-based function prediction methods. A functional relationship between two genes (center) can be inferred from: *A*, chromosomal proximity, both genes are close to each other in a significant number of genomes; *B*, gene fusion/fission, both genes form a single gene in other genome(s); *C*, phylogenetic profiling, both genes have similar *1*) or complementary *2*) phylogenetic profiles in a number of genomes from diverse species; *D*, correlation of gene loss and/or duplication events, the history of losses and duplications of both families is similar; *E*, correlated mutations, pairs of positions, one from each family's protein alignment, vary in the same species; this might indicate that these positions physically interact. In other words, if mutations in position X of protein A are found to be significantly associated (they co-occur in many species), with mutations in position Y of protein B, the prediction can be made that proteins A and B physically interact and that residues Y and X are involved in this interaction. In the example, residues 5, 67, 212, and 310 of the first alignment mutate in the same species as residues 124, 330, 32, and 159, respectively, from the second alignment; *F*, correlation of their performance in genome-wide experiments, especially when this correlation is conserved between different experiments and/or different species; *G*, similarity of the phylogenetic trees of both protein families; this can be measured, as in the example, by the correlation coefficient of the distance matrices from both trees.

conservation of shared performance in genome-wide experiments.

In the late nineties it was shown that the conservation of chromosomal proximity of two genes (Fig. 2*A*) might be indicative of related function (15, 47). This is especially true for bacterial genomes in which genes coding for enzymes in a common pathway are often organized in operons. In eukaryotes, the proximity of two genes in the genome is rarely indicative of related function, although there are several remarkable examples, such as the polycistronic transcripts in nematodes (7). Therefore these techniques are especially suited to mitochondrial proteins that have bacterial homologs, assuming that the functional relationships are conserved between the eukaryotic and bacterial counterparts. A special case of chromosomal proximity is that of fused genes (Fig. 2*B*). The finding of a fusion between two genes in another genome is usually a strong indication for a physical interaction between the proteins encoded therein (38).

Regardless of the proximity in the chromosome, being encoded in the same genome can be a prerequisite for functional interaction. Thus the finding of two genes that co-occur in many genomes and are missed from many others suggests that they likely participate in the same biological process (34, 68, 71). This technique, called gene co-occurrence or phylogenetic profiling (Fig. 2*C*) compares the patterns of presence/absence of proteins in a set of complete genomes and infers

functional interactions between proteins with similar profiles (50). Refinements of this approach (Fig. 2*D*) include the use of the evolutionary relationships among the species considered to identify correlated gene loss or duplication events (23). Other variants that use the coevolution of proteins to predict their function exploit the information contained in the sequences or in the phylogenetic trees. For instance, phylogenies from interacting protein families, such as the chemokine-receptor system (29), are more similar to each other than expected from the species phylogeny. Such correlated evolution can be detected by comparing pairs of trees (63) (Fig. 2*G*) or detecting compensatory mutations from the protein alignments that would indicate a possible protein-protein interaction (49) (Fig. 2*E*).

The third type of genomic context, consists of the performance of genes in genome-wide experiments (Fig. 2*F*). So far, gene expression studies and large-scale protein-protein interaction screenings are the types of experimental genomic context that have been most widely used (64, 65). The inherent noisy nature of this kind of data can be reduced by searching for conservation of the shared genomic context for a pair of genes. This can be done by comparing different experiments in a single species (vertical comparative genomics) or by comparing results of similar experiments in different species (horizontal comparative genomics). Although still far from the overall popularity of homology-based function prediction methods, context-based methods are slowly becoming a general tool for researchers, thanks to user-friendly servers such as STRING (66).

CLINICAL RELEVANCE OF CONTEXT-BASED FUNCTION PREDICTION: SOME SUCCESSFUL CASE-STORIES

Although still in their infancy, context-based function prediction methods have already produced a decent number of successful examples that underscore their validity and applicability. Here I will describe a few case stories, that besides illustrating the mechanics of context-based function prediction in the mitochondrion, have been applied to disease-related processes and thus are of clinical importance. A pioneering case of context-based function prediction applied to a mitochondrial disease is that of the functional annotation of the frataxin gene, in which the phylogenetic profiling technique was used (Fig. 2*C*). For many years, a triplet expansion in an intron of this gene was known to cause the human degenerative disease Friedreich ataxia (10). This slowly progressive disorder of the nervous and muscular systems is caused by a degeneration of nerve tissue in the spinal chord and nerves extending to peripheral areas, such as the arms and legs. With time, this damage leads to the inability to coordinate voluntary muscle movements, without deterioration of mental capacity. Although the mutation causing the disease was well characterized, the molecular function of the frataxin gene and the mechanism by which the mutation was causing the disease remained elusive. By applying the phylogenetic profiling technique, Huynen and coworkers (35) noted that the frataxin gene displayed the same phylogenetic distribution as several other genes involved in the assembly of iron-sulfur clusters in the mitochondrion. Frataxin is present in all eukaryotic genomes analyzed, but its pattern of presence/absence among bacterial genomes is quite distinctive. For instance, it was present in the alpha-proteobacterium *R. prowazekii*, but absent from the closely related *Mesorhizobium loti* and *Caulobacter crescentus*. Exactly the same pattern in all genomes studied was found for *hscA* and *hscB* genes, known to be involved in the assembly of the iron-sulfur cluster and suggesting a similar role for frataxin. This hint inspired researchers in the field and, only a year later, several independent experimental results that confirmed a role of frataxin in the iron-sulfur cluster assembly pathway were published (11, 16, 42).

The frataxin case constitutes an example of how one can predict the function of an uncharacterized protein by identifying the biological process in which it is playing a role. Conversely, genomic-context information can also be exploited when no candidate genes are available but, instead, the goal is to identify them. Such reverse strategy was applied by Gabaldón and colleagues in their search for genes involved in complex I deficiency (27), using this time the correlation of gene gain and loss technique (Fig. 2*D*). Complex I deficiency consists of a reduced activity in the mitochondrial respiratory chain enzyme NADH:ubiquinone oxidoreductase (complex I). Such an impairment may be present in a variety of forms and often results in multisystem disorders associated with a fatal outcome at a young age (62). The intricate macromolecular structure of complex I, comprising 46 subunits (8), complicates the task of identifying the disease-causing gene, since mutations in almost any of its subunits can, in principle, result in complex I deficiency (56). In the worst scenario, after sequencing all known complex I subunits in patients with hereditary complex I deficiency, the disease-causing mutation may remain unidentified. This suggests that the mutation possibly lies in a gene coding for an unknown subunit or a protein directly or indirectly involved in complex I function (e.g., an assembly factor). To help with identifying such genes, Gabaldón and colleagues performed a large-scale phylogenetic analysis involving all subunits of complex I (27). Additionally, they searched for proteins with a gene loss profile similar to that of core complex I subunits (23). One of the candidate genes identified, coding for an uncharacterized protein, showed a number of characteristics that were indicative of a tight functional link with complex I. First, the gene itself (later named B17.2-like or B17.2L) is homologous to the known complex I accessory subunit B17.2. Second, the evolutionary reconstruction of this family reveals that, after a gene duplication event occurred at the early stages of eukaryotic evolution, both paralogous genes evolved in a similar fashion. Most remarkably, both genes have been concomitantly lost from species that also lack complex I such as the fungi *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Encephalitozoon cuniculi*. Such a pattern of coevolution strongly suggests a similar role for B17.2L in complex I function. Only an accelerated evolution of this paralogous group, indicated by the long branches in the tree, suggested a possible subfunctionalization. The experimental confirmation that B17.2L is indeed involved in complex I function came soon, when this gene was found to be mutated in patients with a complex I deficiency associated with progressive encephalopathy (46). The molecular characterization of the protein encoded by B17.2L showed that it was not a permanent component of complex I, but was rather functioning as a chaperone in its assembly process (46).

In some other cases, the experimentalists do have a list of candidate genes that might be causing the disease, but this set is just too long and sequencing or testing all candidates becomes infeasible. In such cases, computational approaches can be used to prioritize the candidates and thus allow not only a faster identification of the disease-causing mutation, but also a more rational use of available resources. All context-based techniques can be used in isolation, but a higher specificity is expected if a combination of methods is used. Initial integrative analysis for discovering genes involved in mitochondrion-associated diseases combined few sources of data to prioritize genes found in a homozygosity region known to cause the disease. For instance, Mootha and colleagues (41) identified the gene causing cytochrome *c* oxidase deficiency by combining proteomics and a large-scale gene expression data set. For all genes encoded in the candidate regions they evaluated they likelihood of being mitochondrial by analyzing their profiles in subcellular proteomics and comparing their expression profiles with those of known mitochondrial proteins. Among the 15 genes encoded in the candidate region, only the mRNA binding protein LRPPRC was found to be clearly associated with mitochondria. The sequencing of that gene in patient samples identified the mutation causing the disease. Later, the same method served to identify mutations in the mitochondrial *ETHE1* gene as the origin of ethylmalonic encephalopathy (61). More recently, integrative approaches have incorporated a growing number of data sources to identify disease genes in candidate regions. This allows applying these methods to larger lists of genes. An example of such a case is the identification of a genomic rearrangement in the succinyl-CoA synthase (*SUCLA2*) gene as the origin of a severe encephalomyopathy (18). A genome-wide homozygosity screen, in a family with several members affected by an autosomal recessive encephalomyopathy allowed the identification of a shared homozygosity region of 20 Mb on chromosome 13. This region contains 103 open reading frames, a list just too long for a comprehensive experimental testing. Since the disease was specifically associated with a mtDNA depletion and a reduced activity in several mitochondrial respiratory complexes, the researchers decided to prioritize the candidate genes according to their possible mitochondrial localization. For this purpose, they employed the MitoP2 score and reduced the original list to only those three candidate genes that presented a MitoP2 score higher than 60. Subsequent experimental characterization of their sequences identified the disease-causing genomic rearrangement in the *SUCLA2* gene. A similar approach, but this time using the *Maestro* Bayesian classifier (9), identified the mitochondrial inner membrane protein MPV17 as the disease gene associated with an infantile hepatic mitochondrial DNA depletion (58).

Concluding Remarks

To provide a very brief summary of the current state of the art regarding the functional characterization of the mitochondrion it can be said that *1*) we are quickly approaching the full identification of the mitochondrial proteome; *2*) besides some well-studied pathways, our understanding of the functional interactions that occur within the mitochondrion is still rather incomplete. The systematic functional characterization of mitochondrial proteins and the elucidation of their functional interactions will ultimately pave the way to the reconstruction of an in silico model for the human mitochondrion, which constitutes the ultimate goal of mitochondrial systems biology. Such a model could be used to predict the dynamics and patterns of response of the mitochondrion under certain conditions, to test theoretical assumptions, and even to perform virtual experiments that would be otherwise infeasible. We are witnessing the first serious attempts to produce reliable computational models of the mitochondrion, an emerging field that is the focus of other articles in the present journal issue. If we compare the present situation with those of twenty, ten and five years ago, there is no doubt that the functional characterization of the mitochondrion has experienced a significant acceleration. Computational biology has played a major role in this process, and it is likely that, its relative contribution to our understanding of the mitochondrion will only grow in the coming years.

ACKNOWLEDGMENTS

T. Gabaldón is the recipient of a long-term fellowship from EMBO (ALTF-2005–204). The author thanks Martijn A. Huynen, Jan Smeitink, and members of their groups for inspiring discussions in this issue.

REFERENCES

1. **Abe Y, Shodai T, Muto T, Mihara K, Torii H, Nishikawa S, Endo T, and Kohda D.** Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. *Cell* 100: 551–560, 2000.
2. **Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ.** Basic local alignment search tool. *J Mol Biol* 215: 403–410, 1990.
3. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ.** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402, 1997.
4. **Basu S, Bremer E, Zhou C, and Bogenhagen DF.** MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation. *Bioinformatics* 22: 485–492, 2006.
5. **Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, and Eddy SR.** The Pfam protein families database. *Nucleic Acids Res* 32 (Database issue): D138–D141, 2004.
6. **Blake JA, Eppig JT, Bult CJ, Kadin JA, and Richardson JE.** The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res* 34 (Database issue): D562–D567, 2006.
7. **Blumenthal T.** Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* 20: 480–487, 1998.
8. **Brandt U.** Energy converting NADH:quinone oxidoreductase (complex I). *Annu Rev Biochem* 75: 69–92, 2006.
9. **Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, and Mootha VK.** Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* 38: 576–582, 2006.
10. **Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Canizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, De Michele G, Filla A, De Frutos R, Palau F, Patel PI, Di Donato S, Mandel JL, Cocozza S, Koenig M, and Pandolfo M.** Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271: 1423–1427, 1996.
11. **Chen OS, Hemenway S, and Kaplan J.** Inhibition of Fe-S cluster biosynthesis decreases mitochondrial iron export: evidence that Yfh1p affects Fe-S cluster synthesis. *Proc Natl Acad Sci USA* 99: 12321–12326, 2002.
12. **Claros MG.** MitoProt, a Macintosh application for studying mitochondrial proteins. *Comput Appl Biosci* 11: 441–447, 1995.
13. **Cotter D, Guda P, Fahy E, and Subramaniam S.** MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res* 32 (Database issue): D463–D467, 2004.
15. **Dandekar T, Snel B, Huynen M, and Bork P.** Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324–328, 1998.

16. **Duby G, Foury F, Ramazzotti A, Herrmann J, and Lutz T.** A non-essential function for yeast frataxin in iron-sulfur cluster assembly. *Hum Mol Genet* 11: 2635–2643, 2002.
17. **Eddy SR.** Profile hidden Markov models. *Bioinformatics* 14: 755–763, 1998.
18. **Elpeleg O, Miller C, Hershkovitz E, Bitner-Glindzicz M, Bondi-Rubinstein G, Rahman S, Pagnamenta A, Eshhar S, and Saada A.** Deficiency of the ADP-forming succinyl-CoA synthase activity is associated with encephalomyopathy and mitochondrial DNA depletion. *Am J Hum Genet* 76: 1081–1086, 2005.
19. **Emanuelsson O, Nielsen H, Brunak S, and von Heijne G.** Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300: 1005–1016, 2000.
20. **Fitch WM.** Homology a personal view on some of the problems. *Trends Genet* 16: 227–231, 2000.
21. **Forner F, Foster LJ, Campanaro S, Valle G, and Mann M.** Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* 5: 608–619, 2006.
22. **Gabaldón T.** Evolution of proteins and proteomes, a phylogenetics approach. *Evol Bioinformatics Online* 1: 51–56, 2005.
23. **Gabaldón T and Huynen MA.** Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. *Bioinformatics* 21, *Suppl* 2: ii144–ii150, 2005.
24. **Gabaldón T and Huynen MA.** Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* 61: 930–944, 2004.
25. **Gabaldón T and Huynen MA.** Reconstruction of the proto-mitochondrial metabolism. *Science* 301: 609, 2003.
26. **Gabaldón T and Huynen MA.** Shaping the mitochondrial proteome. *Biochim Biophys Acta* 1659: 212–220, 2004.
27. **Gabaldón T, Rainey D, and Huynen MA.** Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (complex I). *J Mol Biol* 348: 857–870, 2005.
28. **Gene Ontology Consortium.** The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 34 (Database issue): D322–D326, 2006.
29. **Goh CS, Bogan AA, Joachimiak M, Walther D, and Cohen FE.** Co-evolution of proteins with their interaction partners. *J Mol Biol* 299: 283–293, 2000.
30. **Gordon DM, Dancis A, and Pain D.** Mechanisms of mitochondrial protein import. *Essays Biochem* 36: 61–73, 2000.
31. **Guda C, Fahy E, and Subramaniam S.** MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* 20: 1785–1794, 2004.
32. **Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, and O'Shea EK.** Global analysis of protein localization in budding yeast. *Nature* 425: 686–691, 2003.
33. **Huynen MA and Bork P.** Measuring genome evolution. *Proc Natl Acad Sci USA* 95: 5849–5856, 1998.
34. **Huynen MA and Snel B.** Gene and context: integrative approaches to genome analysis. *Adv Protein Chem* 54: 345–379, 2000.
35. **Huynen MA, Snel B, Bork P, and Gibson TJ.** The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. *Hum Mol Genet* 10: 2463–2468, 2001.
36. **Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, and Bork P.** SMART4.0: towards genomic data integration. *Nucleic Acids Res* 32 (Database issue): D142–D144, 2004.
37. **Lopez MF, Kristal BS, Chernokalskaya E, Lazarev A, Shestopalov AI, Bogdanova A, and Robinson M.** High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* 21: 3427–3440, 2000.
38. **Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, and Eisenberg D.** Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751–753, 1999.
39. **Marcotte EM, Xenarios I, van Der Bliek AM, and Eisenberg D.** Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci USA* 97: 12115–12120, 2000.
40. **Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M, Patterson N, Lander ES, and Mann M.** Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 115: 629–640, 2003.
41. **Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, and Lander ES.** Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci USA* 100: 605–610, 2003.
42. **Muhlenhoff U, Richhardt N, Ristow M, Kispal G, and Lill R.** The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins. *Hum Mol Genet* 11: 2025–2036, 2002.
43. **Nakai K and Horton P.** PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 24: 34–36, 1999.
44. **Nakai K and Kanehisa M.** A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14: 897–911, 1992.
45. **O'Brien KP, Remm M, and Sonnhammer EL.** Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 33 (Database issue): D476–D480, 2005.
46. **Ogilvie I, Kennaway NG, and Shoubridge EA.** A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. *J Clin Invest* 115: 2784–2792, 2005.
47. **Overbeek R, Fonstein M, D'Souza M, Pusch GD, and Maltsev N.** The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 96: 2896–2901, 1999.
48. **Ozawa T, Sako Y, Sato M, Kitamura T, and Umezawa Y.** A genetic approach to identifying mitochondrial proteins. *Nat Biotechnol* 21: 287–293, 2003.
49. **Pazos F, Helmer-Citterich M, Ausiello G, and Valencia A.** Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 271: 511–523, 1997.
50. **Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, and Yeates TO.** Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96: 4285–4288, 1999.
51. **Prokisch H, Andreoli C, Ahting U, Heiss K, Ruepp A, Scharfe C, and Meitinger T.** MitoP2: the mitochondrial proteome database–now including mouse data. *Nucleic Acids Res* 34 (Database issue): D705–D711, 2006.
52. **Prokisch H, Scharfe C, Camp DG, 2nd Xiao W, David L, Andreoli C, Monroe ME, Moore RJ, Gritsenko MA, Kozany C, Hixson KK, Mottaz HM, Zischka H, Ueffing M, Herman ZS, Davis RW, Meitinger T, Oefner PJ, Smith RD, and Steinmetz LM.** Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol* 2: E160, 2004.
53. **Rost B, Liu J, Nair R, Wrzeszczynski KO, and Ofran Y.** Automatic prediction of protein function. *Cell Mol Life Sci* 60: 2637–2650, 2003.
54. **Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, and Meisinger C.** The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc Natl Acad Sci USA* 100: 13207–13212, 2003.
55. **Small I, Peeters N, Legeai F, and Lurin C.** Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4: 1581–1590, 2004.
56. **Smeitink J, Sengers R, Trijbels F, and van den Heuvel L.** Human NADH:ubiquinone oxidoreductase. *J Bioenerg Biomembr* 33: 259–266, 2001.
57. **Smith TF and Waterman MS.** Identification of common molecular subsequences. *J Mol Biol* 147: 195–197, 1981.
58. **Spinazzola A, Viscomi C, Fernandez-Vizarra E, Carrara F, D'Adamo P, Calvo S, Marsano RM, Donnini C, Weiher H, Strisciuglio P, Parini R, Sarzi E, Chan A, Dimauro S, Rotig A, Gasparini P, Ferrero I, Mootha VK, Tiranti V, and Zeviani M.** MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nat Genet* 38: 570–575, 2006.
59. **Tatusov RL, Fedorova ND, Jackson JJ, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, and Natale DA.** The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41, 2003.
60. **Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, Wiley S, Murphy AN, Gaucher SP, Capaldi RA, Gibson BW, and Ghosh SS.** Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* 21: 281–286, 2003.
61. **Tiranti V, D'Adamo P, Briem E, Ferrari G, Mineri R, Lamantea E, Mandel H, Balestri P, Garcia-Silva MT, Vollmer B, Rinaldo P, Hahn SH, Leonard J, Rahman S, Dionisi-Vici C, Garavaglia B, Gasparini P, and Zeviani M.** Ethylmalonic encephalopathy is caused by mutations in ETHE1, a gene encoding a mitochondrial matrix protein. *Am J Hum Genet* 74: 239–252, 2004.

62. **Triepels RH, Van Den Heuvel LP, Trijbels JM, and Smeitink JA.** Respiratory chain complex I deficiency. *Am J Med Genet* 106: 37–45, 2001.
63. **Valencia A and Pazos F.** Prediction of protein-protein interactions from evolutionary information. *Methods Biochem Anal* 44: 411–426, 2003.
64. **Van Noort V, Snel B, and Huynen MA.** Predicting gene function by conserved co-expression. *Trends Genet* 19: 238–242, 2003.
65. **Vazquez A, Flammini A, Maritan A, and Vespignani A.** Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol* 21: 697–700, 2003.
66. **Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, and Snel B.** STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31: 258–261, 2003.
67. **Voos W, Martin H, Krimmer T, and Pfanner N.** Mechanisms of protein translocation into mitochondria. *Biochim Biophys Acta* 1422: 235–254, 1999.
68. **Wu J, Kasif S, and DeLisi C.** Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* 19: 1524–1530, 2003.
69. **Yang ZR.** Biological applications of support vector machines. *Brief Bioinform* 5: 328–338, 2004.
70. **Yates JR, 3rd, Gilchrist A, Howell KE, and Bergeron JJ.** Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol* 6: 702–714, 2005.
71. **Zheng Y, Roberts RJ, and Kasif S.** Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol* 3: RESEARCH0060, 2002.