BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 4 2006, pages 485–492
doi:10.1093/bioinformatics/btk009

Databases and ontologies

MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation

Siddhartha Basu, Erich Bremer, Chun Zhou and Daniel F. Bogenhagen*

Department of Pharmacological Sciences, State University of New York at Stony Brook, Stony Brook, NY 11794-8651, USA

Received on December 5, 2005; accepted on December 13, 2005
Advance Access publication December 20, 2005
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: There has been an explosion of interest in the role of mitochondria in programmed cell death and other fundamental pathological processes underlying the development of human diseases. Nevertheless, the inventory of mitochondrial proteins encoded in the nuclear genome remains incomplete, providing an impediment to mitochondrial research at the interface with systems biology. We created the MiGenes database to further define the scope of the mitochondrial proteome in humans and model organisms including mice, rats, flies and worms as well as budding and fission yeasts. MiGenes is intended to stimulate mitochondrial research using model organisms.

Summary: MiGenes is a large-scale relational database that is automatically updated to keep pace with advances in mitochondrial proteomics and is curated to assure that the designation of proteins as mitochondrial reflects gene ontology (GO) annotations supported by high-quality evidence codes. A set of postulates is proposed to help define which proteins are authentic components of mitochondria. MiGenes incorporates >1160 new GO annotations to human, mouse and rat protein records, 370 of which represent the first GO annotation reflecting a mitochondrial localization. MiGenes employs a flexible search interface that permits batchwise accession number searches to support high-throughput proteomic studies. A web interface is provided to permit members of the mitochondrial research community to suggest modifications in protein annotations or mitochondrial status.

Availability: MiGenes is available at http://www.pharm.stonybrook.edu/migenes

Supplementary Information: Supplementary data are available at Bioinfo online.

Contact: dan@pharm.sunysb.edu

INTRODUCTION

Mitochondria are important organelles in higher eukaryotes well-known for their role in cellular ATP synthesis by oxidative phosphorylation coupled with electron transport. Mitochondria perform crucial steps for heme biosynthesis, calcium signaling, amino acid metabolism, synthesis of Fe/S clusters, and metabolism of urea, amino acids, sterols and vitamin D as well as generation of cellular reactive oxygen species (ROS). Mitochondria maintain scavenging enzymes to protect the cell from ROS in addition to a regulated inventory of pro- and anti-apoptotic factors involved in programmed cell death. The mitochondrion is the only organelle aside from the nucleus that contains its own genomic DNA, requiring separate machinery for its maintenance and expression. Given the prominence of mitochondria in diverse cellular processes, it is not surprising that mitochondrial dysfunction in humans contributes to a multitude of problems, including cancer, cardiomyopathy, diabetes, obesity and many types of neuronal degenerative diseases (Wallace, 2005).

Mitochondrial DNA encodes only a handful of proteins, 13 in vertebrates, requiring the vast majority of mitochondrial proteins to be nuclear-encoded. Though the exact number is still unknown, studies have suggested that mammalian mitochondria may contain as many as 1500 proteins (Taylor *et al*., 2003b). Several recent proteomic surveys have tried to address this issue by identifying the complete set of mitochondrial proteins. Studies with mitochondria from yeast (Sickmann *et al*., 2003) and human heart (Taylor *et al*., 2003a) identified ~750 and ~544 proteins, respectively. A recent study of mouse mitochondria generated a list of 591 proteins and emphasized the conclusion that the mitochondrial proteome varies in a tissue-specific manner (Mootha, 2003). In addition to proteomics, several other genetic, genomic and bioinformatic approaches have also been employed to increase the repertoire of mitochondrial proteins. However, despite these efforts, a substantial fraction of the mitochondrial proteome still remains unidentified. The goal of determining the complete set of mitochondrial proteins is becoming far more complicated and unmanageable as the datasets expand and, to an increasing extent, include false positive identifications. Many of these false positives represent proteins sequenced as contaminants in partially purified mitochondrial fractions. The annotation of some of these proteins as mitochondrial is a result of a lack of defined standards that is confusing to researchers seeking to understand the roles of newly discovered mitochondrial proteins. These difficulties are compounded by the current state of bioinformatics in which a protein may be known and cited under a variety of name and accession number aliases.

Several databases have been generated to collate mitochondrial proteins, including MitoProteome (http://www.mitoproteome.org/) (Cotter *et al*., 2004), the Human Mitochondrial Protein Database established by the National Institute of Standards and Technology (http://bioinfo.nist.gov:8080/examples/servlets/index.html) and MITOP2 (http://ihg.gsf.de/mitop2) (Andreoli *et al*., 2004). All of these databases have their own strengths and weaknesses. For example, MitoProteome and NIST are limited to human records, so that their ability to incorporate results obtained with model organisms

*To whom correspondence should be addressed.

is limited. MITOP2 provides an excellent resource for the mitochondrial research community, but shares some technical problems with the other two databases. For example, MITOP2 cannot be searched with GenBank accession numbers. These databases are not particularly intuitive to the user; they often lack support for flexible searches using various accession numbers and the search interfaces sometimes require prior knowledge of database format or content to retrieve information. The absence of sufficient curation and heavy reliance on electronic annotation also permit the inclusion of inappropriate or invalidated mitochondrial proteins.

Here we report the development of a mitochondrial protein database 'MiGenes' in which we collected and integrated datasets from a number of public sources that encompass mitochondrial proteins of seven eukaryotic organisms. At this time, the database does not include non-protein entries, such as RNAs that may be imported into mitochondria. MiGenes supports a variety of accession number-based as well as text-based keyword searches through a single flexible interface. A batchwise retrieval mode has been tailored for high-throughput accession number-based searches. Every record in MiGenes, with few exceptions, is represented by a curated 'RefSeq' accession number and is mapped to an appropriate 'UniProt' accession number. We developed a 'GO' based search procedure and are in the process of curating records to establish an evidence-based roster of mitochondrial proteins (Ashburner *et al*., 2000). The MiGenes database is designed as a tool to help collate evidence that qualifies novel proteins as mitochondrial.

MATERIALS AND METHODS

Database design and data handling strategies

MiGenes is designed with 14 relational database tables that are conceptually divided into gene and protein sections. Each entry in the database is composed of a connecting gene and protein. For entries with multiple protein isoforms, a gene is linked to multiple proteins. Both gene and protein entries are populated with annotation, sequence identifiers and accession numbers as obtained from the public datasets. Each section can function independently and is capable of receiving different sources of annotation. For each section we have chosen a unique identifier: 'Entrez gene id' for genes and 'RefSeq NP_ accession number' for proteins. We linked GO annotations to the protein section to avoid instances of multiple isoforms receiving the same GO terms. Redundancies are avoided by disallowing the inclusion of shared accession numbers and sequence identifiers. Every entry in the database is assigned an initial status of 'mitochondrial,' 'mitochondria-related' or 'unrelated' based on the associated GO terms and evidence codes. Entries without any GO annotation of mitochondrial location or function are classified 'unrelated.' The initial status of an entry can only be altered by an authorized curator in a process that requires adding or negating a particular combination of GO term and evidence code. If GO annotations that change the status of a protein are added by our database sources, the status is automatically adjusted unless the change would violate other curator-annotated decisions regarding status.

Data pipeline

The MiGenes database is populated by an automated process that involves harvesting and integrating datasets from disparate public databases (Fig. 1). These files are then processed by a series of upload programs roughly equivalent to the steps of MiGenes build processes. For each build step and for each database section, the linking of accession numbers precedes the uploading of annotation datasets. The building process starts by populating the gene section with Entrezgene (ftp.ncbi.nlm.nih.gov/gene/DATA/) and UniGene (ftp.ncbi.nih.gov/repository/UniGene/) datasets and then, for every gene, creates a placeholder for corresponding protein sections. Next, the protein placeholders are populated with accession numbers and annotations from the RefSeq (ftp.ncbi.nih.gov/refseq/release/) datasets. UniProt accession numbers and annotations (www.pir.UniProt.org/database/download.shtml) are then added to the protein section from its mapping with RefSeq accession numbers. The mappings of RefSeq and UniProt accession numbers, however, are prepared through a separate build process described below. Next, from the non-redundant dataset (ftp.ncbi.nlm.nih.gov/blast/db/), another set of accession numbers and sequence identifiers is extracted and uploaded to the protein section. Lastly, the protein section is populated with homolog information from NCBI homologene (ftp.ncbi.nih.gov/pub/HomoloGene/current) and GO annotations (ftp.ebi.ac.uk/pub/databases/GO/goa/ and ftp.geneontology.org/pub/go/gene-associations/) from the gene-association datasets. MiGenes is updated frequently in synchrony with the release of the latest public datasets. However, instead of rebuilding the database at every interval, the update process is run only for the stages for which the latest release becomes available. No accession number or sequence identifier is removed during updates; they are maintained in linkage with the current identifier, thereby providing better compatibility with publications that refer to older accession numbers. When a primary RefSeq identifier is replaced, for example when an XP number is replaced by an NP number, the replaced identifier is retained as an alias.

Validated mapping of Refseq and UniProt accession numbers

To expand the cross-indexing or mapping of accession numbers between major pubic databases, we first extracted the mapping data available from various public datasets. International Protein Index (IPI; ftp.ebi.ac.uk/pub/databases/IPI/current; human, mouse and rat; Kersey *et al*., 2004), GenBank non-redundant datasets (nr; ftp.ncbi.nlm.nih.gov/blast/db/; all organism), Mouse Genome Informatics (MGI; ftp.informatics.jax.org/pub/reports/MRK_Sequence.rpt and MRK_SwissProt_TrEMBL.rpt; Blake *et al*., 2003) and *Saccharomyces* Genome Database (genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/dbxref.tab; yeast) datasets were processed with bl2seq (pairwise BLAST) to generate the initial dataset of mapped pairs. To generate a uniform, integrated dataset, we performed an intra-organism reciprocal BLASTP search with RefSeq and UniProt datasets. Both the public and our in-house mapped pairs were then merged and the identical pairs were grouped into single entries. We employed a conservative cutoff for a mapped pair where sequence identity is >90%, difference of length is <10 amino acids and the cumulative gap length is <5% of the total alignment length. Of the 3481 mapped pairs identified by this procedure 3431 mapped pairs were identical in length and sequence, but permitting variability allowed us to evaluate and link the remaining 50 pairs that had slight discrepancies in the GenBank and UniProt data sources. UniProt numbers for every pair that score above this cutoff are designated as primary links with the corresponding RefSeq numbers.

Design and construction of the GO browser

The organization of the MiGenes GO browser was created from the DAG (directed acyclic graph) structure of GO terms. However, the content of the graph was modified to include only the terms and relationships that are relevant to mitochondria. From the three ontologies, we selected less-specialized terms that broadly account for the established aspects of mitochondrial biology. From each of the selected terms, we collected groups of subordinate terms that describe specific aspects of mitochondrial functions (Supplementary Table 1). Both the broad and specific terms are used for classification of status, whereas only the broad terms are used for display. If present, secondary identifiers and synonyms of relevant GO terms are also included. In effect, the MiGenes GO browser presents higher level less-specific terms where the specific child terms remain implicit. GO terms and relationships are downloaded as mysql dumps (http://archive.godatabase.org/latest-full/) and then uploaded to the MiGenes table. The tree structure

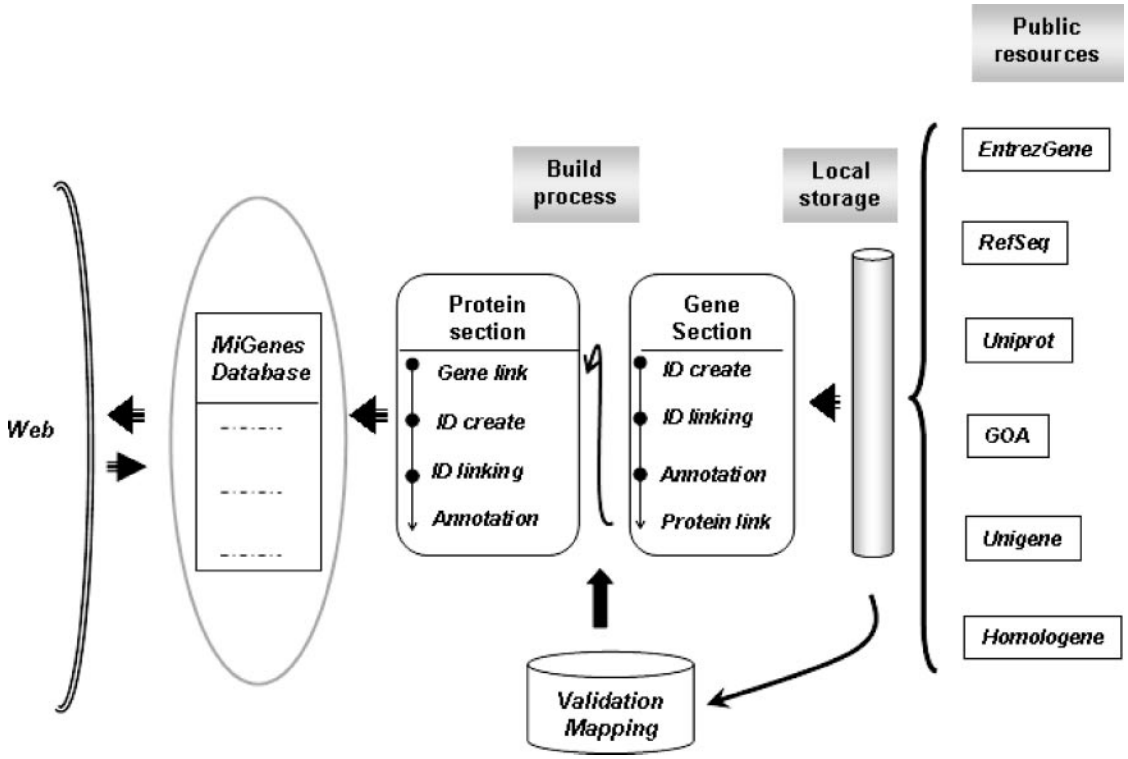

Fig. 1. MiGenes dataflow. Data processing starts with the downloading of data from public sources on the right, followed by the application of algorithms to coordinate and validate mapping of records between diverse sources and to sort records according to their mitochondrial status. A variety of display options make the database accessible via a web interface. Lists of records matching a query can be downloaded as a tab-delimited list to be imported into a spreadsheet such as Excel.

is first built from the four tables of the GO database schema: term, term2-term, term-synonym and graph path. The structure is then modified dynamically by applying the filtering criteria from an internal MiGenes table which keeps track of terms that are used to display the tree and gene-association.

Software implementation

In MiGenes, harvesting of public datasets, building of validated mapping and upload process are performed by a series of Perl scripts. Bioperl libraries (Stajich *et al*., 2002) are used extensively for parsing, transformation and local access of biological data records. GO-perl libraries (http://www.godatabase.org/dev/index.html) are used for parsing of gene ontology gene-association files. The web interface is implemented in Sun Java (JDK1.4- http://java.sun.com) using java servlet technology running on an Apache Tomcat 5.5.9 (http://jakarta.apache.org/tomcat/index.html) server. A Mysql 4.1× (http://www.mysql.com) server is used as the storage engine for the database

RESULTS AND DISCUSSION

The problem-oriented design of MiGenes. Problem 1: identifying and cross-referencing records in diverse data sources

In our efforts to organize information on mitochondrial proteomes, we encountered a number of problems that we endeavored to solve within the MiGenes database, some of which were pointed out in the Introduction. One of our goals was to merge data from important mammalian and non-mammalian species in a single database in order to take advantage of the wealth of information available from these model organisms that may support or refute the conclusion that a particular protein plays a role in mitochondria. We found that data resources on mitochondrial proteins were scattered; in many cases genomic sequencing projects did not carefully consider the subcellular location of proteins. The exception to this was the yeast, *Saccharomyces cerevisiae*, where extensive data analysis has provided a substantial data resource (Prokisch *et al*., 2004; Sickmann *et al*., 2003).

One of the major obstacles we encountered in building MiGenes was that we observed that many GenBank entries were not cross-indexed, or mapped, to UniProt entries, or vice versa. In addition, many of the entries that were cross-indexed did not have identical protein sequences in the two databases. Some specialized databases used their own identifiers that were typically not mapped to both GenBank and UniProt entries. The data flow and build process used to coordinate data from diverse sources in order to construct the MiGenes database is described in Materials and Methods and illustrated in Figure 1. We adopted the convention that proteins should be identified by their NCBI RefSeq numbers whenever possible and that UniProt, other sequence identifiers (GI) and specialized database identifiers should be cross-referenced as accession number aliases. In the process of linking records, we performed systematic BLAST searches to align and validate the mapping (or cross-indexing) between Refseq and UniProt accession numbers. We imported UniProt based gene-associations for worm, fly and fission yeast that are not available from the RefSeq dataset. This enables users to identify the correct RefSeq entity using any valid accession

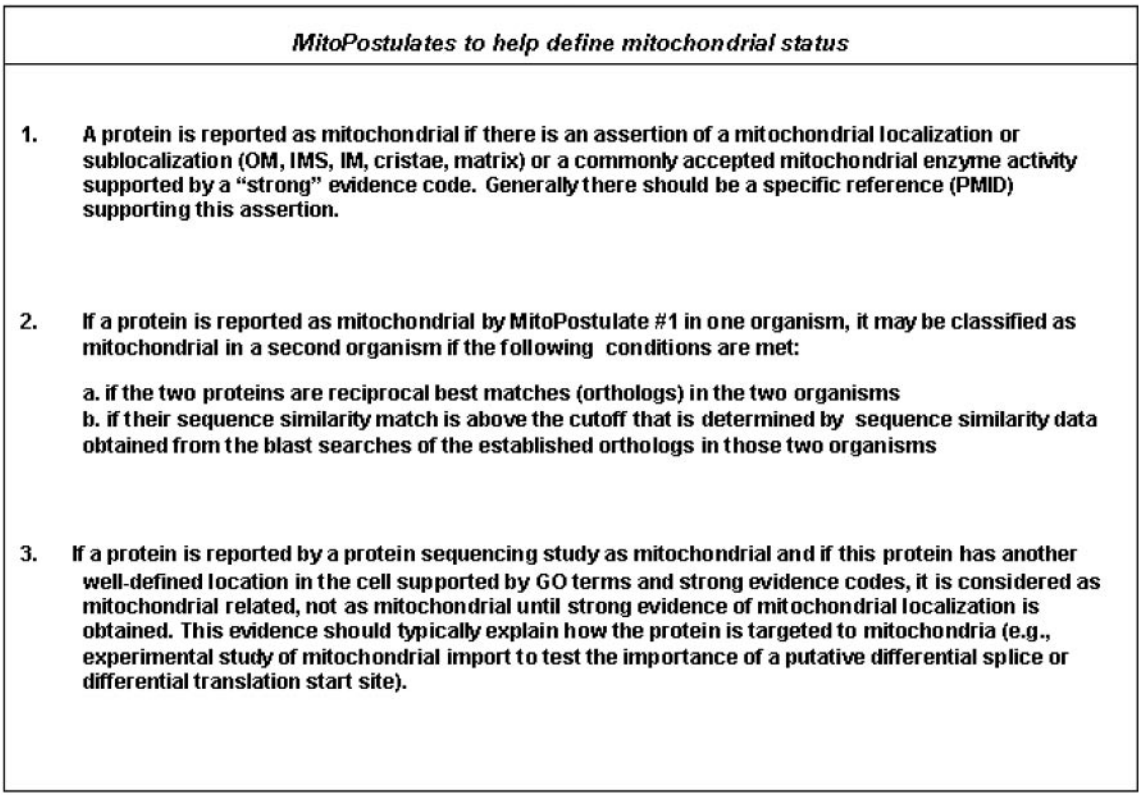

Fig. 2. Mitopostulates for classification of proteins as mitochondrial.

number alias. This goal was particularly important for interpretation of protein lists in high-throughput sequencing studies. For example, a large study of the mouse mitochondrial proteome by Mootha *et al*., (2003) reported results using GI sequence identifiers, whereas another study by Da Cruz *et al*., (2003) reported UniProt accession numbers of inner membrane mitochondrial proteins. The cross-reference mappings of the MiGenes project allowed us to integrate and analyze these diverse datasets.

We encountered instances of accession numbers reported in published datasets that have been replaced by new identifiers in the public databases. We incorporated a historical tracking mechanism in the MiGenes database that permits the current record to be identified even by retired accession numbers. However, maintaining a comprehensive list of accession number aliases remains a challenging problem and we still occasionally encounter instances where cited accession numbers fail to identify any records. We estimate that, on average, a typical MiGenes entry can be accessed through over 150 accession numbers and sequence identifiers.

Problem 2: determining which proteins are truly mitochondrial

A protein is considered mitochondrial if solid evidence supports GO annotations that reflect either location in a mitochondrial subcompartment, molecular functions explicitly defined as mitochondrial or a role in biological processes widely accepted as occurring only in mitochondria (terms shown in Supplementary Table 1). These GO terms permit some uncertainty. For example, while we accept cellular component annotations for mitochondrial matrix, inner and outer membrane, and intermembrane space, we tried not to include a protein that simply adhered to the outer membrane unless the protein was an intrinsic membrane protein, like monoamine oxidase. We adopted criteria that we refer to as 'Mito Postulates' to help decide whether a protein is mitochondrial (Fig. 2). The first guideline stipulates that the mitochondrial localization or function of a protein must be supported by an appropriate GO descriptor supported by one or more of the strong evidence codes listed in Table 1. Other GO terms were not considered convincing, such as NAS or 'non-traceable author statement,' a type of evidence that

Table 1. List of strong and weak GO evidence codes used to classify records in MiGenes

Strong evidence codes	Weak evidence codes
IMP, inferred from mutant phenotype	ISS, inferred from sequence similarity
IGI, inferred from genetic interaction	IEA, inferred from electronic annotation
IPI, inferred from physical interaction	IEP, inferred from expression pattern
TAS, traceable author statement	NAS, non-traceable author statement
IC, inferred by curator	ND, no biological data available
IDA, inferred from direct assay	RCA, Inferred from reviewed computational analysis

might be considered hearsay in a judicial context. In contrast, we accept the evidence code TAS, 'traceable author statement' when accompanied by a literature citation (PubMed ID) in the absence of contradictory evidence of high quality. Other questionable evidence codes are ISS, 'inferred by sequence similarity' and IEP 'inferred from expression pattern.' The validity of the ISS code clearly depends on the criteria used as a cutoff, as discussed in more detail later. We considered IEP to be a weak criterion specifically to avoid gene products from being automatically classified as mitochondrial based on an expression pattern similar to known mitochondrial proteins. Setting these strict criteria required that we reject some proteins from the mitochondrial group pending further investigation. Proteins tagged by GO terms supported by less definitive evidence codes were classified as 'mitochondria-related' proteins. This is admittedly a broad 'catch all' category intended to identify proteins to track to determine whether new evidence might support a mitochondrial designation in the future. Listing a protein as 'mitochondria-related' may direct further studies to test its mitochondrial localization.

These criteria provide only the initial assessment of the potential mitochondrial role of a protein. Many mitochondrial proteins may be missed if the existing GO annotations are not complete or accurate. Owing to errors in the GO database, our criteria permitted entry into the MiGenes database of a number of proteins that we considered false positives. The single greatest source of these false positives was found to be electronic annotation of records as mitochondrial using the evidence code IDA to reflect the fact that the protein had been identified when proteins in a mitochondrial cell fraction were sequenced using high-throughput or shotgun approaches. Given the power of modern mass spectrometric protein sequencing, it is physically impossible to purify mitochondria with no contaminating proteins from other compartments. We adopted the standard of placing a protein in the 'mitochondria-related' category when an IDA evidence code based on protein sequencing was in conflict with an established role in another cellular compartment.

A large number of proteins are annotated with conflicting evidence codes relating to their subcellular localization, generating considerable confusion. It is appropriate to permit a protein record to carry apparently conflicting location codes, since many genes are expressed in alternative forms with different cellular locations. In many cases this reflects the existence of distinct protein isoforms targeted at different cellular locations, as discussed later. We have not been able to generate a fully automated procedure to deal with such situations. Our solution has been to manually curate records in the 'mitochondria-related' category to review the evidence for and against mitochondrial localization.

During the process of developing and testing these criteria, we noticed that the liberal use of electronic annotations has introduced a very large number of GO term assignments that are either incorrect or inconsistent with other GO terms. For example, high-throughput gene discovery efforts have given many mitochondrial proteins a GO assignment reflecting an extracellular location. Since the GO database is currently used as a very common resource for annotation of gene and protein function, every effort must be made to improve the accuracy of GO term assignments.

Curation of functional protein groups Automated sorting of human entries revealed a large number of apparently authentic mitochondrial proteins with 'mitochondria-related' status. We have adjusted the mitochondrial status of many proteins by manual curation from literature references. For example, the initial annotation of mammalian mitochondrial ribosomal proteins using only automated procedures provided only a short list of mitochondrial ribosomal proteins. We analyzed data from Suzuki *et al*. (2001a,b) and Koc *et al*. (2001a,b) and assigned 'mitochondrial' status to 61 additional human and 31 additional mouse ribosomal proteins. From the work of Vo *et al*. (2004) we annotated 11 proteins involved in mitochondrial metabolism and promoted them to 'mitochondrial' status. A total of 10 mitochondrial carrier proteins were also annotated as mitochondrial based on the reviews of Palmieri, (2004) and del Arco *et al*. (2004).

Orthology-based curation Sequence similarity is widely used as a criterion to assign a function to a conserved protein. However, especially for newly-sequenced genomes, where many genes may be predicted by gene-recognition software, this homology-based criterion has the potential to link a gene whose function is well known in one organism to a paralog or pseudogene of its true ortholog in another organism. Our second Mitopostulate (Fig. 2) expresses clear guidelines for the conservative use of sequence homology as a criterion for assigning mitochondrial status. During our curation process, we generated a list of 284 human–mouse orthologs (from the NCBI homologene project) where both proteins had mitochondrial status. BLAST searches showed that 95% of these protein pairs differed in length by <20 residues and shared at least 70% sequence identity. The NCBI homologene dataset included an additional 176 human–mouse protein pairs where one ortholog had been proven to be mitochondrial but the other had 'mitochondria-related' status. By applying the cutoffs established for true orthologs, we resolved nearly 80% of these discrepancies in mitochondrial status. The classification of rat entries was accomplished by comparing them with both human and mouse 'mitochondrial' orthologs, permitting 270 rat entries to be promoted to 'mitochondrial' status (Fig. 3). In cases where the BLASTP search for rat proteins related to a known human and mouse mitochondrial protein identified a best reciprocal match that did not meet our cutoff criteria, this rat protein would be left in the 'mitochondria-related' category.

We next extended this orthology-based curation to human–mouse protein pairs where the status of one of the proteins was initially 'unrelated' to mitochondrial function, indicating that one protein lacked any electronic annotation of mitochondrial status. This search identified proteins that were given 'mitochondrial' status by automated GO annotations even when the protein had no obvious association with mitochondria but did have conflicting GO term assignments for another cellular compartment. A total of 73 mouse proteins were demoted to 'unrelated' status following this analysis, including obvious errors of high-throughput sequencing such as annotation of trypsinogen and hemoglobin as mitochondrial. Additional mouse entries, 15, were demoted from 'mitochondrial' to 'mitochondria-related' as they warrant further experimental attention to confirm their association.

In the case of invertebrates, worm and fly, the majority of the entries categorized as 'mitochondria-related' were tagged with weak evidence codes such as ISS or IEA. Lack of extensive experimental support coupled with their evolutionary distances from their mammalian partners required us to adopt a conservative approach during our initial curation efforts. Entries with at least two

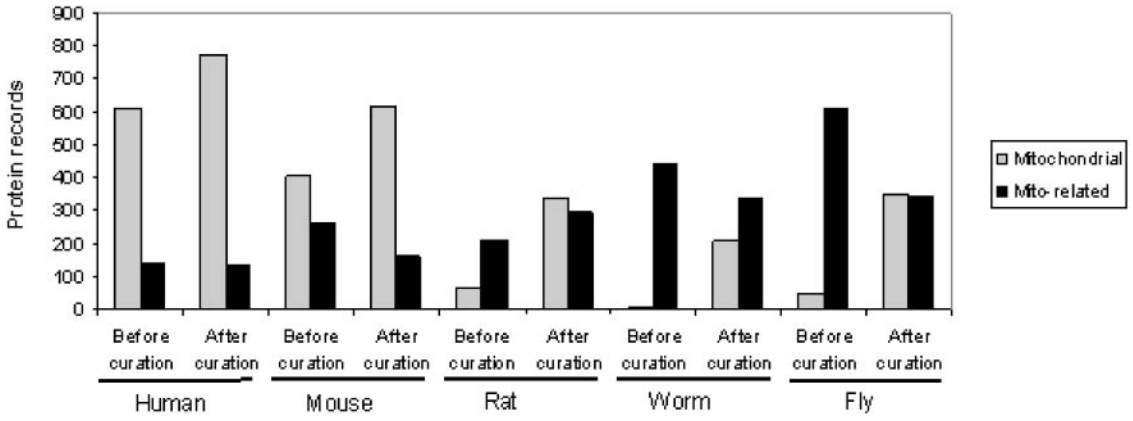

Fig. 3. Effect of orthology-based curation on the MiGenes protein records of various organisms. Mitopostulate #2 was applied for inter-species comparison of orthologs that resulted in the reclassification of the mitochondrial status of numerous records. The resultant lists were then used for another round of comparison and curation as a control for the manual curation process.

Table 2. Enumeration of MiGenes records (as of 10.5.2005)

Organism	Genes		Proteins	
	Mitochondrial	Mitochondria-related	Mitochondrial	Mitochondria-related
Homo sapiens	594	179	717	221
Mus musculus	606	159	614	163
Rattus norvegicus	386	271	386	271
Drosophila melanogaster	382	233	400	324
Caenorhabditis elegans	224	326	224	334
Saccharomyces cerevisiae	994	41	994	41
Saccharomyces pombe	36	484	36	484

mitochondrial mammalian orthologs having pairwise identity not <40% and size difference not >20 amino acids were considered for curation. For the curation of fly entries, the dataset of Tripoli *et al*. (2005) was employed as additional support material. This resulted in promotion of 298 fly and 207 worm entries to 'mitochondrial' status (Fig. 3).

Overall, 1162 mammalian and 574 fly or worm protein entries were curated with a new mitochondrial GO term using our 'Mitopostulates.' These include 370 mammalian entries (96 human, 126 mouse and 148 rat) that were annotated with a mitochondrial GO term for the first time. Table 2 shows the numbers of records classified as 'mitochondrial' and 'mitochondria-related' in the seven reference organisms. Supplementary Table 2 lists over 1700 GO annotation recommendations sent to the GO consortium following this initial round of curation.

Problem 3: dealing with multiple protein isoforms linked to a single gene locus

One of the greatest obstacles facing proteomic research is the problem of dealing with multiple isoforms of a protein encoded by a single genetic locus. Many mitochondrial proteins are isoforms of proteins found in other cellular locations (Danpure, 1995). Differential localization is often achieved by differential translation initiation or differential splicing. The multiplicity of protein isoforms complicates protein identification using mass spectrometry since different isoforms can give different ion signatures. In many cases, the amount of protein sequence information obtained in an experiment is not sufficient to distinguish which of several isoforms may be present in a particular sample since the identified peptides may be common to several isoforms. Thus, protein sequencing studies often do not resolve the issue of defining which isoforms of a protein may be imported into mitochondria. When a single isoform is shown to be mitochondrial, the public database often adds to the complexity by annotating the gene locus, thereby designating all the isoforms as mitochondrial. In addition, NCBI RefSeq and UniProt records sometimes designate different isoforms as the canonical sequence, so that these are sometimes not the best reciprocal matches. It is a long-term goal of the MiGenes database to help resolve inconsistent annotations of mitochondrial proteins represented by multiple isoforms. This issue will only be resolved by careful annotation that is beyond the scope of the initial release of MiGenes. We consider that MiGenes will be a very useful tool for continued annotation of mitochondrial proteins.

Description of MiGenes Functions

Search interfaces The basic search interface (Fig. 4) is highly flexible to anticipate any search string a researcher is likely to use. It can be queried with text, keywords, accession numbers or sequence

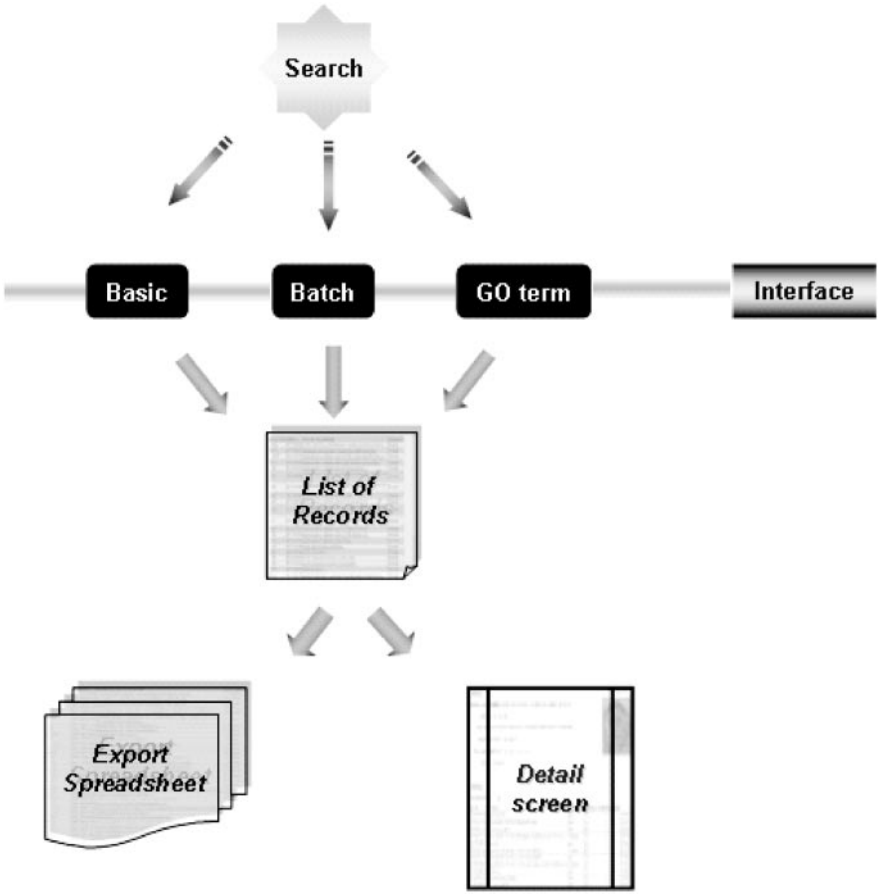

Fig. 4. MiGenes search strategies. All three interfaces generate lists of records matching search criteria. The user can then either download the list in a spreadsheet format or navigate further to the detailed report screen for an individual record of interest.

identifiers. By default, it is restricted to a single organism selected from a drop-down list. A wide range of known and publicly available accession number and sequence identifiers are supported and can be queried through this interface. In the simplest case, a unique matching record is returned. For an ambiguous search, a list of records may be returned where the isoforms, if present are grouped together. Every record in the list is displayed as a hyperlink to its detail report screen.

The batch search interface can be queried with a list of accession numbers or sequence identifiers that can include a mixture of heterogeneous accession numbers or sequence identifiers. The report of matching records can be exported as an Excel spreadsheet. This makes the batch search suitable for post-processing of high-throughput protein sequencing data that generates a list of accession numbers or sequence identifiers as its final output.

The GO search interface is built around those GO terms that are appropriate for describing the functions of mitochondrial gene products. Our GO browser displays a trimmed-down version of the directed acyclic graph structure (Supplementary Table 1). A search through this interface retrieves lists of proteins annotated with the selected GO terms. Modifications of GO term assignments can be made by our curators through a curator interface, which is not publicly accessible. However, users can submit recommendations for modifications to GO terms through a comment screen that requests the term to be changed, the supporting evidence code and an applicable PubMed identifier. It is our intention to collect recommendations for changes in the GO term assignments suggested by our curators and users of the database to be communicated regularly to the GO consortium. This ensures a proper dissemination of high quality curated data to the entire research community, which is one of the major goals of the MiGenes database.

Data Displays MiGenes has not been designed as a primary data repository, but rather as a gateway to collate information available from centralized sequence databases. The main single-record report screen is organized into two sections for a better understanding and analysis of annotations that are available in MiGenes. The first section of the record describes the name of gene and protein along with all the aliases associated with it. Primary sequences and other information are accessed through links to Entrez gene (Maglott *et al*., 2005), RefSeq (Pruitt *et al*., 2005) and UniProt (Bairoch *et al*., 2005) public databases. In order to keep most pertinent data on one screen, a drop-down selection is provided to access any of four expanded lists of relevant information. These options include (1) a list of published references, (2) a list of orthologous proteins in other organisms, (3) comments generated

by curators and (4) a list of GO annotations describing properties of the protein. These GO annotations and their evidence codes, with PubMed ID's, provide the basis for classification of a protein as 'mitochondrial' or 'mitochondria-related.'

ACKNOWLEDGEMENTS

This work was supported by grant 5 R01 ES012039 from NIEHS to D.F.B.

Conflict of Interest: none declared.

REFERENCES

Andreoli,C. *et al*. (2004) MitoP2, an integrated database on mitochondrial proteins in yeast and man. *Nucleic Acids Res*., **32**, D459–D462.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Blake,J.A. *et al*. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res*., **31**, 193–195.

Cotter,D. *et al*. (2004) MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res*., **32**, D463–D467.

Da Cruz,S. *et al*. (2003) Proteomic analysis of the mouse liver mitochondrial inner membrane. *J. Biol. Chem*., **278**, 41566–41571.

Danpure,C.J. (1995) How can the products of a single gene be localized to more than one intracellular compartment? *Trends Cell. Biol*., **5**, 230–238.

del Arco,A. and Satrustegui,J. (2004) Identification of a novel human subfamily of mitochondrial carriers with calcium-binding domains. *J. Biol. Chem*., **279**, 24701–24713.

Kersey,P.J. *et al*. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.

Koc,E.C. *et al*. (2001a) The small subunit of the mammalian mitochondrial ribosome. Identification of the full complement of ribosomal proteins present. *J. Biol. Chem*., **276**, 19363–19374.

Koc,E.C. *et al*. (2001b) The large subunit of the mammalian mitochondrial ribosome. Analysis of the complement of ribosomal proteins present. *J. Biol. Chem*., **276**, 43958–43969.

Maglott,D. *et al*. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.

Mootha,V.K. *et al*. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, **115**, 629–640.

Palmieri,F. (2004) The mitochondrial transporter family (SLC25): physiological and pathological implications. *Pflugers Arch*., **447**, 689–709.

Prokisch,H. *et al*. (2004) Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol*., **2**, e160.

Pruitt,K.D. *et al*. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.

Sickmann,A. *et al*. (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl Acad. Sci, USA*, **100**, 13207–13212.

Stajich,J.E. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.

Suzuki,T. *et al*. (2001a) Structural compensation for the deficit of rRNA with proteins in the mammalian mitochondrial ribosome: systematic analysis of protein components of the large ribosomal subunit from mammalian mitochondria. *J. Biol. Chem*., **276**, 21724–21736.

Suzuki,T. *et al*. (2001b) Proteomic analysis of the mammalian mitochondrial ribosome. Identification of protein components in the 28 S small subunit. *J. Biol. Chem*., **276**, 33181–33195.

Taylor,S. *et al*. (2003a) Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol*., **21**, 282–286.

Taylor,S.W. *et al*. (2003b) Global organellar proteomics. *Trends Biotechnol*., **21**, 82–88.

Tripoli,G. *et al*. (2005) Comparison of the oxidative phosphorylation (OXPHOS) nuclear genes in the genomes of *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. *Genome Biol*., **6**, R11.

Vo,T.D. *et al*. (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem*., **279**, 39532–39540.

Wallace,D.C. (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet*., **39**, 359–407.