Nucleic Acids Research, 2004, Vol. 32, Database issue ***D459–D462***
DOI: 10.1093/nar/gkh137

MitoP2, an integrated database on mitochondrial proteins in yeast and man

C. Andreoli, H. Prokisch*, K. Hörtnagel, J. C. Mueller, M. Münsterkötter[1], C. Scharfe[2] and T. Meitinger

Institute of Human Genetics and [1]Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany and [2]Department of Biochemistry and Stanford Genome Technology Center, 855 California Avenue, Palo Alto, CA 94304, USA

Received September 17, 2003; Revised and Accepted October 27, 2003

ABSTRACT

The aim of the MitoP2 database (http://ihg.gsf.de/mitop2) is to provide a comprehensive list of mitochondrial proteins of yeast and man. Based on the current literature we created an annotated reference set of yeast and human proteins. In addition, data sets relevant to the study of the mitochondrial proteome are integrated and accessible via search tools and links. They include computational predictions of signalling sequences, and summarize results from proteome mapping, mutant screening, expression profiling, protein–protein interaction and cellular sublocalization studies. For each individual approach, specificity and sensitivity for allocating mitochondrial proteins was calculated. By providing the evidence for mitochondrial candidate proteins the MitoP2 database lends itself to the genetic characterization of human mitochondriopathies.

INTRODUCTION

A complete description of the molecular components of a particular cell type and its subcompartments is an essential prerequisite for any systematic biological approach. The mitochondrion is one of the best studied cellular compartments with a central role in metabolism and energy production and lends itself to such a strategy (1). Current estimates exceed the figure of 1000 for the number of mitochondrial proteins that are encoded almost entirely by nuclear genes. Only a very small fraction of mitochondrial proteins, 12 in yeast and 13 in man, are encoded by the mitochondrial genome. The correlation between sequence changes and functional impairment of mitochondria associated with disease has been initially focused on the organellar genome (2). During the last years, an ever increasing list of nuclear genes has been identified with mutations leading to mitochondrial dysfunction (3).

MITOCHONDRIAL DATABASES

While several databases are established covering the genetic variation of the human mitochondrial genome (4,5), the MitoP2 database (http://ihg.gsf.de/mitop2) focuses on the nuclear-encoded proteome of mitochondria. The annotation of mitochondrial proteins in the generic databases is incomplete and does not always distinguish between proteins which have a confirmed mitochondrial sublocalization and those which are only candidates according to preliminary experimental results or *in silico* predictions. The MitoP2 database tries to fill this gap by integrating the information related to mitochondrial proteins from various resources. Tools have been implemented which allow searches according to various parameters. The MitoP2 database constitutes an update and extends the original Mitop database (6). So far it focuses on the mitochondrial proteome of yeast and man. Extensions to mouse, *Caenorhabditis elegans* and *Arabidopsis thaliana* will be added soon. The aim of the database is to provide a comprehensive reference list of mitochondrial proteins and to link individual entries with databases containing information about the protein components of the organelle.

THE MITOCHONDRIAL PROTEOME OF YEAST

Saccharomyces cerevisiae is the most intensively investigated model organism also with respect to the largest data set available on mitochondrial proteins generated by a wide spectrum of high-throughput experiments. Yeast provides a unique set of mutants and a large number of genetic and molecular tools for functional analysis. A collection of null mutants exists for about 6000 ORFs of the yeast genome. There are extensive descriptions available on pet mutants (7–9). Systematic sublocalization studies have been performed (10). Expression profiles have been published which compare RNA levels of cells under different genetic and environmental conditions (11,12). Yeast also plays a leading role in proteomics (13). Proteomic mapping studies require large amounts of purified samples of the organelle, which can be easily provided by the unicellular organism. Apart from such experimental data, affiliation to the subcellular compartment can be judged according to the presence or absence of mitochondrial signal sequences (MSS) which are used by the import machinery of the organelle to select incoming proteins (14).

The yeast MitoP2 database lists all 6516 ORFs of the yeast genome (Table 1) (19). The list includes a high confidence

*To whom correspondence should be addressed. Tel: +49 89 3187 2890; Fax: +49 89 3187 3297; Email: prokisch@gsf.de

Table 1. Data sources and number of *S.cerevisiae* entries integrated in the MitoP2 database

Data source	Number of entries
ORFs	
SGD April 2003	6516
Mitochondrial reference set	477
Deletion phenotype screening	
pet (8)	381
class III (9)	466
Comprehensive localization studies	
subloc_01 (10)	2746
Protein–protein interactions	
single experiments (CYGD)	6555
high-throughput experiments	78981
Computational predictions	
PSORT (15)	981
MitoProt (score > 0.9) (16)	574
Predotar	397
Bayesian system (17)	500
Bidirectional best-to-best Blast hits (MitoP2)	
yeast–*R.prowazekii*	931
yeast–*E.cuniculi*	5915
yeast–human (total)	1893
yeast–human (mitochondrial reference set)	174
Mitochondrion-bound polysomes (MLR) (18)	3106
Transcriptome	
transc_01 (11)	514
transc_02 (12)	416
Proteomics	
prot_01 (13)	177

reference set of experientially validated mitochondrial proteins encompassing 477 entries. All entries with an annotated mitochondrial localization in Mitop (6), SGD (19) and CYGD (20) were screened manually in the original literature for direct evidence in single experiments. Entries with only indirect evidence were excluded to reduce annotation errors (21). Results from high-throughput experiments are listed separately. The reference set of 477 mitochondrial proteins has been sorted into major functional entities such as 'protein translation/stability' or 'biogenesis of iron–sulfur cluster'. The annotation of these categories is accomplished by a consortium of mitochondria researchers organized in a European network (MitEURO, www.miteuro.org). Using the MitoP2 reference set of known mitochondrial proteins, we assessed the sensitivity and specificity of the different high-throughput approaches. The result is summarized in a single page which can be accessed by clicking the 'general info' button in the search mask (Fig. 1).

The search mask allows a combined search for gene names, keywords and components of the integrated data set. The combination of search criteria provides lists of annotated mitochondrial proteins and new candidates with high confidence. By selecting the subcellular localization experiment by Kumar *et al.* (10) and an *in silico* prediction tool such as Predotar, MitoP2 extracts a list of 141 proteins. It contains 109 proteins from the mitochondrial reference list and 32 new candidates fulfilling the two selected criteria. Each individual protein entry—in a single line—contains the ORF description

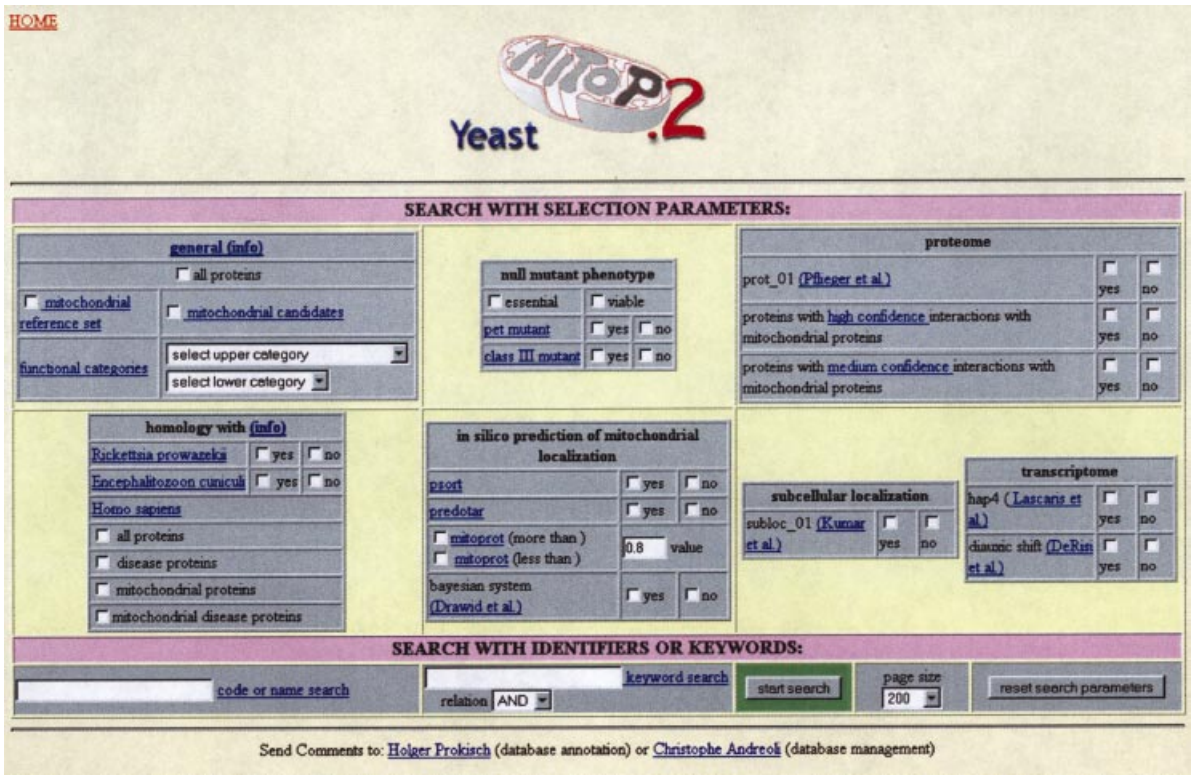

Figure 1. Search mask for mitochondrial proteins in yeast. Datasets can be selected from high-throughput experiments and *in silico* predictions.

Table 2. Data sources and number of human entries integrated in the MitoP2 database

Data source	Number of entries
Human proteins	
Swiss-Prot release 41.9	44496
Mitochondrial reference set	
MitoP2	656
Disease genes	
Swiss-Prot release 41.9	1331
Computational predictions	
PSORT (15)	6125
MitoProt (score > 0.9) (16)	1603
Predotar	908
Bidirectional best-to-best Blast hits (MitoP2)	
human–*R.prowazekii*	1427
human–*E.cuniculi*	38891
human–yeast (total)	1681
human–yeast (mitochondrial reference set)	232
Proteomics	
prot_01 (25)	544

from SGD, the subcellular localization and the information from high-throughput experiments and *in silico* calculations. In addition to the matrix provided by the columns with systematic experimental results and *in silico* data, the user can access detailed information for each protein: an extra page linked to each entry provides (i) a description and localization of the individual protein compiled from SGD and additional genetic and biophysical properties from MIPS data (CYGD), (ii) the corresponding entry from the Gene Ontology (GO) database (22), (iii) a list of homologies to sequences of other species generated by bidirectional best BLAST hits and (iv) a compilation of protein–protein interactions weighted according to von Mering *et al.* (23).

THE MITOCHONDRIAL PROTEOME OF MAN

In contrast to the situation in yeast, the human mitochondrial proteome is about double the size but much less data are available from high-throughput experiments. The current list of human mitochondrial proteins in the MitoP2 database comprises 656 entries and is based on Swiss-Prot entries (24) and manually annotated entries from the original Mitop database (6). This collection constitutes the human mitochondrial reference set.

In order to extend this list, we performed several calculations using genome-wide DNA and protein data sets (Table 2). These included homology searches between human, yeast, *Encephalitozoon cuniculi* [an obligate intracellular parasite lacking mitochondria (26)] and *Rickettsia prowazeckii* [an obligate intracellular parasite closely related to mitochondria (27)]. Additionally we used established algorithms [MitoProt (16), PSORT (15), Predotar] to predict the subcellular localization of a protein based on the amino acid sequence. The proteome analysis of mitochondria purified from heart tissue is so far the only high-throughput experiment available for humans and has been integrated (25). A search mask enables the user to screen these data sets by combined queries (Fig. 2).

The MitoP2 database lists 44 996 human Swiss-Prot entries and—in a single line—presents the description and localization as annotated in Swiss-Prot including the Swiss-Prot link. In addition it summarizes the information from *in silico* calculations and the proteome experiment. Apart from this matrix, the user can access detailed information for each single protein: an extra page for each protein provides (i) a list of homologies to sequences from other species including the Blast E-value and the percentage of the aligned protein length (yeast homologues are linked to the MitoP2 yeast page), (ii) if existent, a short Swiss-Prot description of an associated

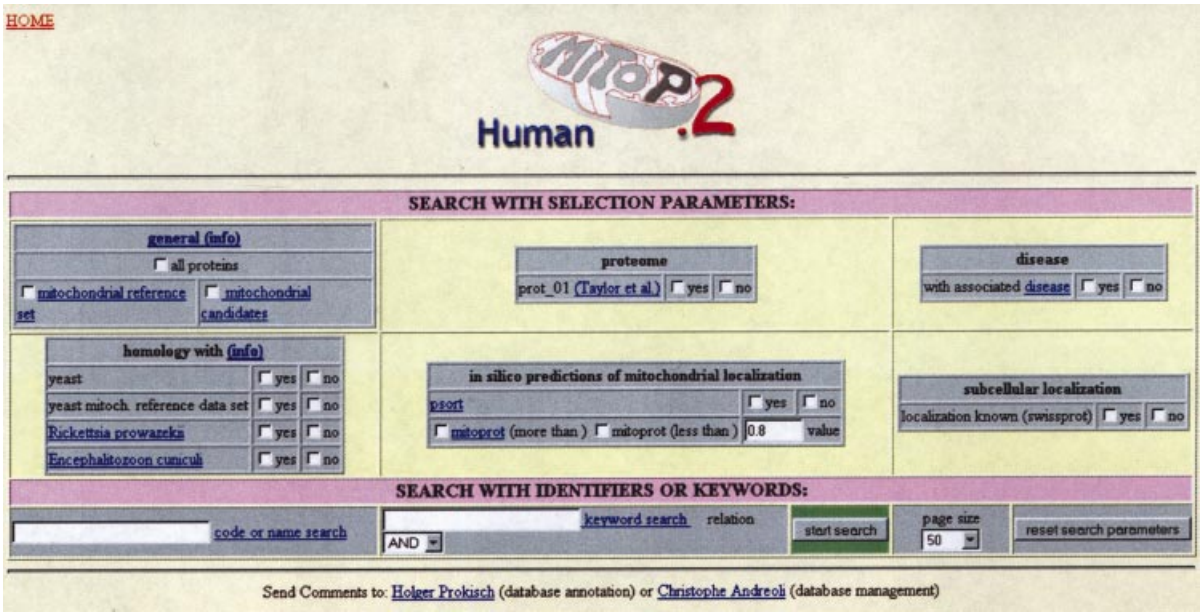

Figure 2. Search mask for mitochondrial proteins in man. Datasets can be selected from high-throughput experiments and *in silico* predictions.

disease [linked to OMIM (28)], (iii) the corresponding GO annotations for molecular protein function, biological process and cellular component (linked to GO), (iv) the available literature about the protein and protein variants listed with authors and title and (v) a table of cross-references annotated in Swiss-Prot.

Based on the current annotations, we found that of 1331 identified human disease proteins, a proportion of ~10% (129) of the known disease proteins are known to be localized in mitochondria. Since the localization of many disease proteins is unknown, this might be an underestimate. The existence of a human mitochondrial homologue for a yeast protein is a very good predictor for mitochondrial localization in yeast. Vice versa, the next to complete set of yeast mitochondrial proteins helped to generate a comprehensive list of 900 human entries with best bidirectional Blast hits to their yeast counterparts. There are more than 300 disease entries in OMIM where the disease locus is unknown and which list clinical signs characteristic of mitochondriopathies. Thus, the human candidate list presents a powerful tool for the genetic characterization of these human mitochondriopathies.

ACKNOWLEDGEMENTS

We wish to thank the MitEURO consortium for their contribution especially Ian Small, Chris Leaver and Lee Sweetlove for their input on mitochondrial proteome data. The Predotar calculation was kindly performed by Ian Small. MitoP2 is funded by the BMBF projects DHGP (German human genome project) and BFAM (Bioinformatics for the Functional Analysis of Mammalian Genomes).

REFERENCES

1. Scheffler,I.E. (2001) Mitochondria make a come back. *Adv. Drug Deliv. Rev.*, **49**, 3–26.
2. DiMauro,S. and Schon,E.A. (1998) Nuclear power and mitochondrial disease. *Nature Genet.*, **19**, 214–215.
3. Zeviani,M., Spinazzola,A. and Carelli,V. (2003) Nuclear genes in mitochondrial disorders. *Curr. Opin. Genet. Dev.*, **13**, 262–270.
4. Kogelnik,A.M., Lott,M.T., Brown,M.D. Navathe,S.B. and Wallace,D.C. (1998) MITOMAP: a human mitochondrial genome database—1998 update. *Nucleic Acids Res.*, **26**, 112–115.
5. Attimonelli,M., Altamura,N., Benne,R., Brennicke,A., Cooper,J.M., D'Elia,D., Montalvo,A., Pinto,B., De Robertis,M., Golik,P. *et al.* (2000) MitBASE: a comprehensive and integrated mitochondrial DNA database. The present status. *Nucleic Acids Res.*, **28**, 148–152.
6. Scharfe,C., Zaccaria,P., Hoertnagel,K., Jaksch,M., Klopstock,T., Dembowski,M., Lill,R., Prokisch,H., Gerbitz,K.D., Neupert,W. *et al.* (2000) MITOP, the mitochondrial proteome database: 2000 update. *Nucleic Acids Res.*, **28**, 155–158.
7. Tzagaloff,A. and Dieckmann,C.L. (1990) PET genes in *Saccharomyces cerevisiae*. *Microbiol. Rev.*, **54**, 211–225.
8. Dimmer,K.S., Fritz,S., Fuchs,F., Messerschmitt,M., Weinbach,N., Neupert,W. and Westermann,B. (2002) Genetic basis of mitochondrial function and morphology in *Saccharomyces cerevisiae*. *Mol. Biol. Cell.*, **13**, 847–853.
9. Steinmetz,L.M., Scharfe,C., Deutschbauer,A.M., Mokranjac,D., Herman,Z.S., Jones,T., Chu,A.M., Giaever,G., Prokisch,H., Oefner,P.J. *et al.* (2002) Systematic screen for human disease genes in yeast. *Nature Genet.*, **31**, 400–404.
10. Kumar,A., Cheung,K.H., Tosches,N., Masiar,P., Liu,Y., Miller,P. and Snyder,M. (2002) The TRIPLES database: a community resource for yeast molecular biology. *Nucleic Acids Res.*, **30**, 73–75.
11. Lascaris,R., Bussemaker,H.J., Boorsma,A., Piper,M., van der Spek,H., Grivell,L. and Blom,J. (2003) Hap4p overexpression in glucose-grown *Saccharomyces cerevisiae* induces cells to enter a novel metabolic state. *Genome Biol.*, **4**, R3.
12. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
13. Pflieger,D., Le Caer,J.P., Lemaire,C., Bernard,B.A. Dujardin,G. and Rossier,J. (2002) Systematic identification of mitochondrial proteins by LC-MS/MS. *Anal. Chem.*, **74**, 2400–2406.
14. Neupert,W. (1997) Protein import into mitochondria. *Annu. Rev. Biochem.*, **66**, 863–917.
15. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
16. Claros,M.G. (1995) MitoProt, a Macintosh application for studying mitochondrial proteins. *Comput. Appl. Biosci.*, **11**, 441–447.
17. Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301**, 1059–1075.
18. Marc,P., Margeot,A., Devaux,F., Blugeon,C., Corral-Debrinski,M. and Jacq,C. (2002) Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO Rep.*, **3**, 159–164.
19. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
20. Mewes,H.W., Frishman,D., Güldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Münsterkoetter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
21. Gilks,W.R., Audit,B., Angelis,D., Tsoka,S. and Ouzounis,C.A. (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.
22. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
23. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
24. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
25. Taylor,S.W., Fahy,E., Zhang,B., Glenn,G.M., Warnock,D.E., Wiley,S., Murphy,A.N., Gaucher,S.P., Capaldi,R.A., Gibson,B.W. *et al.* (2003) Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.*, **21**, 281–286.
26. Katinka,M.D., Duprat,S., Cornillot,E., Metenier,G., Thomarat,F., Prensier,G., Barbe,V., Peyretaillade,E., Brottier,P., Wincker,P. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 401–402.
27. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H. and Kurland,C.G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 109–110.
28. McKusick,V.A. (1998) *Mendelian Inheritance in Man: Catalogs of Human Genes and Genetic Disorders*. Johns Hopkins University Press, Baltimore, MD.