Eur. J. Biochem. *241*, 779–786 (1996)

Computational method to predict mitochondrially imported proteins and their targeting sequences

Manuel G. CLAROS[1,2] and Pierre VINCENS[3]

[1] Laboratoire de Génétique Moléculaire, CNRS URA 1302, Ecole Normale Supérieure, Paris, France
[2] Laboratorio de Bioquímica y Biología Molecular, Facultad de Ciencias, Universidad de Málaga, Spain
[3] Departement de Biologie, Ecole Normale Supérieure, Paris, France

(Received 9 August 1996) – EJB 96 1197/3

Most of the proteins that are used in mitochondria are imported through the double membrane of the organelle. The information that guides the protein to mitochondria is contained in its sequence and structure, although no direct evidence can be obtained. In this article, discriminant analysis has been performed with 47 parameters and a large set of mitochondrial proteins extracted from the SwissProt database. A computational method that facilitates the analysis and objective prediction of mitochondrially imported proteins has been developed. If only the amino acid sequence is considered, 75–97% of the mitochondrial proteins studied have been predicted to be imported into mitochondria. Moreover, the existence of mitochondrial-targeting sequences is predicted in 76–94% of the analyzed mitochondrial precursor proteins. As a practical application, the number of unknown yeast open reading frames that might be mitochondrial proteins has been predicted, which revealed that many of them are clustered.

Keywords: mitochondria; protein import; discriminant function; targeting sequence; transit peptide.

Eukaryotic organelles are mainly composed of proteins synthesized in the cytoplasm. Newly synthesized proteins must be specifically targeted to the correct destination. This requires an efficient, specific addressing system, which has not been determined for all proteins. Mitochondria are organelles that contain their own genetic system. Despite this, most of the hundreds of proteins that they use are encoded by nuclear DNA, synthesized in the cytoplasm and imported. Precursor forms of mitochondrial nuclear-encoded proteins contain targeting and sorting signals that are essential to direct them to mitochondria and their various compartments. Increasing experimental evidence on the nature of these signals has been obtained [1, 2]. They can be present as a cleavable N-terminal signal sequence or within the mature parts of precursor proteins (reviewed in [3]). The most abundant and best studied are the N-terminal presequences. They are extremely variable in length and amino acid sequence, and no evident similarity occurs among them. However, several common properties have been described, such as the enrichment of Arg, Leu, Ser and Ala, the presence of at least two positively charged residues, the paucity of acidic residues, and the ability to form α-helical amphiphilic structures [4–10]. It has been demonstrated that mitochondrial proteins have a high isoelectric point [11] and that their import is subject to hydrophobic constraints [12, 13]. These properties are not enough to target a mitochondrial protein since analyses of the N-terminal regions of cytosolic proteins or randomly generated sequences can emulate the pattern of a targeting sequence [4]. There are many cryptic mitochondrial targeting sequences, as demonstrated in [14], in which about 2.7% of random fragments encoded by total *E. coli* DNA were found to act as mitochondrial-targeting sequences, or in [15], in which an internal amphiphilic helix of the dihydrofolate reductase was able to a direct a hydrophilic protein to the mitochondrial matrix. Therefore, an objective way to predict mitochondrial proteins and their targeting sequences is required.

Theoretical studies of mitochondrial proteins that take into account the characteristics mentioned above have recently been published [16]. That article made evident the difficulties in applying standard biophysical tools to structural analyses, because the data are given without prediction. The number of properties to be fulfilled by mitochondrial proteins is high and there are a lot of exceptions. Therefore, interpretation of the data is necessary, which may lead to controversial conclusions. In this work, methodological approaches were developed, in which the number of parameters used to characterize mitochondrial proteins was enlarged, and the number of proteins studied in relation to previously published articles [4, 6, 16, 17] was increased. This has permitted the development of a computational method that provides an objective means, when only the sequence is available, to determine whether a protein possesses a mitochondrial-targeting sequence, the location of such sequence, and whether a protein could be mitochondrial.

Correspondence to M. G. Claros, Laboratorio de Bioquímica y Biología Molecular, Facultad de Ciencias, Campus de Teatinos, E-29071 Málaga, Spain
Fax: +34 5 213 20 00.
E-mail: claros@uma.es or claros@cica.es

Abbreviations. DFM, discriminant function for mitochondrial proteins; ECS, Eisenberg's consensus scale; GES, Goldman, Engelman and Steitz scale; GvH1, Gunnar von Heijne scale; KD, Kyte and Doolittle scale; LDF, linear discriminant function; MTSL, mitochondrial-targeting-sequence localization; PCA, principal-component analysis.

MATERIALS AND METHODS

Sequences and computed parameters. The studied amino-acid-sequence data were collected from the SwissProt release 31.0 (original release, and new sequences published 23 October 95) [18]. Proteins that lack the N-terminal methionine, are coded

in an organelle, are localized in the chloroplast, or are prokaryotic were not considered. Hence all the sequences studied were nuclearly coded without any post-translational modifications. The parameters computed were as follows.

ZoneTo is the number of N-terminal amino acid residues that can be considered as the maximum length of the putative mitochondrial-targeting sequence. The rationale is that negatively charged residues are mostly absent from that targeting region. Therefore a negative charge followed closely by an additional negative charge can be predicted to lie within the mature part. An Asp or Glu is assumed to belong to the targeting region if it is not followed by another negatively charged residue within 13 residues [7]. ZoneTo is considered up to this point, excluding the first negative charge found.

Coef20 follows the discriminant function for mitochondrial-targeting sequences given in [17]. It indicates whether the abundance of particular residues wihtin the first 20 residues is consistent with a mitochondrial-targeting sequence. For prediction of targeting sequences, the constant in the equation is 0.65, instead of that described, to avoid an excess of false positives (Coef20 > 0).

CoefTot is based on Coef20 taking into account the abundance of more amino acid residues [6] and is calculated for all the region that is supposed to be a targeting sequence, which is delimited by the ZoneTo parameter. The changes in the equation are (a) there is no independent term, (b) the coefficient of R is 0.135 and (c) the numbers of A, S and L are taken into account, with coefficients of 0.141, 0.112 and 0.122, respectively. These changes are an attempt to include the composition calculations described in [16].

CleavSite is the last putative mitochondrial-peptidase-cleavage site in the N-terminal region limited by ZoneTo. Since one of the described rules [7] is not considered, not all the targeting sequences will have a cleavage site predicted. The algorithm to calculate it [16] assigns the cleavage-site position to the residue that will be retained in the mature form.

KR is the number of positively charged residues (Lys and Arg) in the ZoneTo region. Targeting sequences contain, at least, two positively charged residues.

DE is the number of acidic residues (Glu and Asp) in the ZoneTo region. In most cases, no acidic residue is present in a targeting sequence.

ChDiff is the total net charge over the complete amino acid sequence. Mitochondrial proteins are usually positively charged [11].

H17 is the 17-residue segment of higher hydrophobicity in the sequence [12]. Increasing the value diminishes the possibility of importing a protein.

MesoH is the average of the maximal hydrophobicity of a protein over an extended sequence length [12, 16]. It is a way to reflect the proximity between the hydrophobic domains. Hydrophobic domains are not very close to each other in mitochondrially imported proteins.

$\mu H\delta$ is the maximal Eisenberg's hydrophobic moment with δ angles of 75°, 95°, 100° and 105°, with a scanning window of 18 residues [4]. It is applied to the N-terminal region delimited by the cleavage site, or if there is not a large enough amphiphilic segment in this region, by ZoneTo.

Hmaxδ is the maximal hydrophobicity of each hydrophobic face in a helical structure [4]. It is calculated between the 18 residues determined by a maximal $\mu H\delta$.

H17, MesoH, $\mu H\delta$ and Hmaxδ are dependent upon the hydrophobic scale used. To minimize the bias introduced by the scale, calculations have been made with up to four scales based on different amino acid residue properties: Goldman, Engelman and Steitz scale (GES), which reflects the circumstances in which amino acid residues appear in proteins by quantifying the free energy of water/oil transfer for residues in an α-helix structure [19]; Gunnar von Heijne scale 1 (GvH1), which is a statistical scale obtained from the amino-acid-residue frequencies in the central part of a transmembrane segment with respect to the non-membranous stretches [20]; Kyte and Doolittle scale (KD), which takes into account values from water-to-vapor energy transfers and from internal/external distribution of amino acid residues [21]; and Eisenberg's consensus scale (ECS), which was designed to mitigate the effect of outlying values in any one scale, produced by the peculiarities of the method, and is a normalized average of four scales [22].

Multivariate analysis. To study a large data array, the classical approach was to analyze the data by means of principal-component analysis (PCA) [23]. In the present study, proteins were considered as an object defined by a set of variables. Each variable corresponds to measurement of one of 47 parameters derived from the parameters and scales described above (Table 1). Similarly to an object characterized by a unique variable, which can be plotted along a linear axis, each of the above proteins can be mapped onto a 47-dimensional space, even though these variables are correlated. A feature of PCA is that the initial set of n variables describing an object can be replaced by another set in which each new variable is a linear combination of the initial variable, and each new variable is independent and corresponds to a true independent aspect in the classification of the objects. The new axes produced are sorted in such a way that the overall information they contain decreases. As a result of this approach, it is possible to obtain new viewpoints of the data representing the objects by means of software developed in our laboratory [24].

The discriminant analysis computes a function [the linear discriminant function (LDF)] for classification of observations into two or more groups on the basis of n quantitative variables. In our studies, each observation (i.e. the amino acid sequences) is described by a vector of 47 parameters. Assuming that each group has a multivariate normal distribution with different means but equal variance matrices, the computed LDF provides maximum discrimination between groups. The LDF is defined as

$$f(O) = \sum_{i=1}^{n} \alpha_i x_i + C, \qquad (1)$$

where x_i is the value of the variable i for an observation, α_i is the corresponding linear coefficient of the LDF, and C is a constant. The probability of an observation O belonging to group g is equal to

$$P(O/g) = e^{-0.5 f_g(O)} \Big/ \left[\sum_i e^{-0.5 f_i(O)}\right]. \qquad (2)$$

An observation is classified into group u if $g = u$ produces the largest value of $P(O/g)$. A complete description of this data-analysis technique is given in [23]. Discriminant analyses have been performed by means of the package Discrim of the SAS software (SAS Institute Inc.).

Statistical treatments. The jack-knife method [25] was used to validate the results obtained when different training and testing groups of sequences were used.

The statistical significance of results from different computations were carried out as described in [26] with a confidence level of 95%. Values of P indicate the level of significance, consequently putting the lower threshold of significance at $P = 0.05$. The correlation index (R) was calculated as described in [26] and the variables are considered as correlated if $R > 0.5$.

Programs. The computations described in this article have been implemented in a program called MitoProt II. A version

for Macintosh computers, that predicts targeting sequences and the organellar localization, has been written in Symantec THINK Pascal v4.0.2, making extensive use of resources. It has been compiled with MacOS 7.5 operating system. This application is compatible with Macintosh computers running system 6.0.2 or higher. It makes extensive use of the graphic capabilities of Macintosh computers and respects the standard file-handling and window-handling procedures. Texts, tables and default values have been built into resources to allow easy access and ability for permanent modifications by the user. Sequences can be analyzed one by one or in groups of up to 17, in several standard formats. Results can be saved, copied or printed out. All the representations that were available in older Macintosh MitoProt [16], can be displayed to facilitate the interpretation of data, and have been updated to include the new calculations. A UNIX version of Mitoprot II has been written in ADA95, by means of GNAT compiler, on a Sun Sparc Station under System V release 4.0 (Solaris 2.4), without any use of graphic capabilities to allow its use even from computers with dumb terminal settings. This release provides the parameters described above to predict the cell localization, giving the $P(O/g)$ for one given sequence. The results are saved in a file for further manipulation.

Both compiled programs can be freely obtained by anonymous FTP from various servers such as the EMBL (ftp.ebi.ac.uk) or ENS (ftp.ens.fr), upon request to the authors, or on floppy disk (enclose a formatted disk and a self-addressed envelope).

RESULTS AND DISCUSSION

Parameter validity. The analyses of all the sequences of the SwissProt database after application of the restrictions indicated in Materials and Methods classified 12432 sequences as non-mitochondrial and 607 as mitochondrial. Reviewing all the published properties of mitochondrial-targeting sequences, 11 parameters were computed. Because four hydrophobic scales are used to compute H17 and MesoH, as well as four hydrophobic moments ($\mu H\delta$) and hydrophobic faces (Hmaxδ), a total of 47 parameters are used for each sequence (Table 1). These parameters have been validated over a large number of sequences corroborating preliminary results that used a small set of proteins and only one or two hydrophobic scales. The mean values of these parameters agree with those published [4, 5, 11, 12, 27]. However, there is no statistical difference ($P < 0.13$) for acidic residues between mitochondrial and non-mitochondrial sequences, which is logical since, by definition, the studied region is devoid of acidic residues. Furthermore, the mean length of the region that can support a targeting sequence can be applied to 40 residues and is bigger than that for sequences that do not have N-terminal mitochondrial-targeting sequences.

The H17 and MesoH are smaller in mitochondrial proteins than in others, although the difference is only significant with GES ($P < 0.01$). This is not surprising, since most of the proteins in the cell are globular proteins, which will have low H17 and MesoH values. By means of ECS, the mean $\mu H\delta$ and Hmaxδ values for mitochondrial proteins conform with those reported for targeting sequences ($\mu H\delta > 7.3$, Hmax$\delta > 4.4$ [4]), which will be very useful for further studies. The set of variables can be grouped into six main correlated groups: the first includes H17 and MesoH, the second Coef20 and Hmaxδ, the third CoefTot, CleavSite, DE and ZoneTo, the fourth $\mu H75$ with the four scales, the fifth $\mu H95$, $\mu H100$ and $\mu H105$, and the sixth includes ChDiff alone ($R = 0.23$ at best). KR is unusual since it is related to the third, fourth and fifth groups. The correlations observed for the same parameter with different hydrophobic scales is

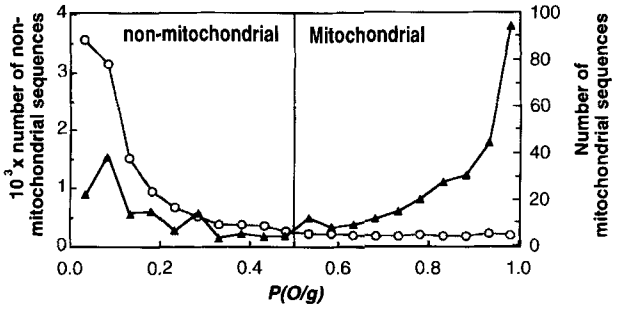

Fig. 1. Distribution of the 13039 proteins from the SwissProt database by means of DFM. $P(O/g) > 0.5$ indicates predicted mitochondrial localization and $P(O/g) < 0.5$ indicates non-mitochondrial localization. The peak of $P(O/g) = 0.2$ for the mitochondrial group is mainly composed by proteins that do not have very clear mitochondrial characteristics (see text for details). ○, non-mitochondrial sequences; ▲, mitochondrial sequences.

significant ($R \approx 0.75$). Although a new qualifier that gathers the values obtained with the different scales could be obtained, in this work the different scales have been retained since this will enable comparison with other published results. In conclusion, the parameters chosen to predict mitochondrial-targeting sequences and mitochondrial localization are significantly different between the two groups and strengthen the results obtained in previous studies, which analyzed a smaller group of proteins. This suggests that these parameters can be used to distinguish between mitochondrially imported proteins and others.

Discriminant analysis for mitochondrial proteins. In the first approach, PCA was performed. A representation of the factorial planes shows that mitochondrial proteins stayed together (data not shown). This result implies that a linear transformation was adapted to separate proteins. Therefore, a discriminant analysis similar to that conducted previously [17, 28], was used to determine optimal function. This analysis has the advantage that it provides a linear function starting with a set of interdependent variables since most of the parameters that are used in this work are correlated.

With the SwissProt database as the training group, the coefficients for the discriminant function that identifies the mitochondrial proteins (DFM), which uses the 47 parameters, are presented in Table 1. Other discriminant functions were assayed with less parameters and less hydrophobicity scales, but the final success rates were never better than 68.80% (data not shown). Fig. 1 shows the distribution of the 13039 sequences according to the probability of belonging to each group. The less numerous group of probability lies near the boundary value of $P(O/g) = 0.5$. Separation of the proteins into those with $P(O/g) > 0.5$ and those with $P(O/g) < 0.5$ as mitochondrial and non-mitochondrial proteins, respectively, showed that DFM provided very accurate approaches for distinguishing between mitochondrial proteins and others (Table 2). The 119 mislocalized mitochondrial proteins were analyzed, which showed that only a few proteins were really mislocalized: 31.9% corresponded to matrix and/or internal-membrane proteins; 24.6% were outer-membrane proteins; 15.1% were without clear localization; and 28.5% did not present transit peptides. This reduced the number of mislocalized proteins to 38 (31.9% of the 119 putative mislocalized proteins, 6.9% of the total protein) that escape predictions, since mitochondrial-outer-membrane proteins and proteins without transit peptides are expected to defy our prediction rules. This

Table 1. Statistics for the 47 parameters that can be used to analyze mitochondrially imported proteins. The mean values are expressed as a confidence interval with a confidence level of 95%; overlapping intervals for the same parameter in the mitochondrial and non-mitochondrial populations indicate no statistical differences. The coefficients are *ai* for the equation DFM. The parameter weight is the product of the mean and coefficient.

Parameter	Mean values for		Coefficients for		Parameter weight for	
	mitochondrial proteins	non-mitochondrial proteins	mitochondrial proteins	non-mitochondrial proteins	mitochondrial proteins	non-mitochondrial proteins
Coef20	4.50 ± 0.04	3.94 ± 0.011	22.02985	20.86752	99.134	82.218
ChDiff	4.03 ± 0.71	−4.04 ± 0.313	0.03617	0.02389	0.146	− 0.097
ZoneTo	40.26 ± 1.72	25.02 ± 0.560	− 0.11407	− 0.11403	− 4.592	− 2.853
KR	5.95 ± 0.27	2.17 ± 0.055	− 0.12308	− 0.32858	− 0.732	− 0.713
DE	0.42 ± 0.06	0.37 ± 0.013	− 1.36920	− 0.07230	− 0.575	− 0.027
CoefTot	−0.08 ± 0.10	−0.47 ± 0.040	− 1.41757	− 1.62884	0.113	0.766
CleavSite	27.50 ± 1.67	9.37 ± 0.425	0.02416	0.01063	0.664	0.100
H17_GES	1.23 ± 0.05	1.43 ± 0.013	− 6.05997	− 5.77033	− 7.454	− 8.252
MesoH_GES	−0.53 ± 0.06	−0.45 ± 0.018	4.51905	3.97155	− 2.395	− 1.787
μH75_GES	31.07 ± 0.91	20.91 ± 0.171	0.33657	0.33495	10.457	7.004
μH95_GES	33.80 ± 0.87	21.25 ± 0.182	0.43619	0.36786	14.743	7.817
μH100_GES	32.39 ± 0.82	21.01 ± 0.183	− 0.18110	− 0.14372	− 5.866	− 3.020
μH105_GES	32.15 ± 0.83	21.06 ± 0.181	0.49205	0.45354	15.819	9.552
Hmax75_GES	10.92 ± 0.47	9.46 ± 0.145	− 0.17395	− 0.21127	− 1.896	− 1.999
Hmax95_GES	11.49 ± 0.42	9.25 ± 0.141	− 0.27610	− 0.29360	− 3.172	− 2.716
Hmax100_GES	11.62 ± 0.42	9.83 ± 0.137	− 0.31709	− 0.28662	− 3.685	− 2.817
Hmax105_GES	11.22 ± 0.44	9.51 ± 0.140	− 0.23148	− 0.25556	− 2.597	− 2.430
H17_KD	1.42 ± 0.03	1.72 ± 0.013	−16.62201	−13.15266	−23.603	−22.623
MesoH_KD	0.23 ± 0.05	0.25 ± 0.017	5.79386	4.00750	1.333	1.002
μH75_KD	19.17 ± 0.54	13.66 ± 0.109	0.48397	0.43010	9.278	5.875
μH95_KD	20.74 ± 0.51	13.89 ± 0.112	0.52823	0.42838	10.955	5.950
μH100_KD	19.77 ± 0.49	13.69 ± 0.111	0.28395	0.34401	5.614	4.710
μH105_KD	19.52 ± 0.49	13.73 ± 0.112	0.09246	0.04840	1.805	0.665
Hmax75_KD	13.24 ± 0.46	11.90 ± 0.142	− 0.43463	− 0.41635	− 5.755	− 4.955
Hmax95_KD	13.51 ± 0.41	11.42 ± 0.135	− 0.72477	− 0.69331	− 9.792	− 7.918
Hmax100_KD	13.97 ± 0.40	11.94 ± 0.134	− 0.38540	− 0.41858	− 5.384	− 4.998
Hmax105_KD	13.44 ± 0.41	11.91 ± 0.138	− 0.46126	− 0.42639	− 6.199	− 5.078
H17_GvH1	0.14 ± 0.01	0.22 ± 0.004	9.11772	6.93752	1.276	1.526
MesoH_GvH1	−0.41 ± 0.04	−0.39 ± 0.014	−72.09837	−66.10239	29.560	25.846
μH75_GvH1	8.16 ± 0.23	5.85 ± 0.047	0.63266	0.66065	5.163	3.865
μH95_GvH1	8.74 ± 0.22	5.96 ± 0.050	1.14993	1.28801	10.050	7.677
μH100_GvH1	8.41 ± 0.21	5.87 ± 0.049	− 1.02931	− 0.84230	− 8.656	− 4.944
μH105_GvH1	8.44 ± 0.21	5.88 ± 0.049	0.70274	0.62314	5.931	3.664
Hmax75_GvH1	1.76 ± 0.16	1.34 ± 0.050	− 0.73638	− 0.64103	− 1.296	− 0.859
Hmax95_GvH1	1.93 ± 0.15	1.24 ± 0.049	− 0.33848	− 0.38797	− 0.653	− 0.481
Hmax100_GvH1	1.98 ± 0.14	1.39 ± 0.048	0.26099	0.23765	0.517	0.330
Hmax105_GvH1	1.98 ± 0.14	1.37 ± 0.049	0.36197	0.24277	0.717	0.333
H17_ECS	0.56 ± 0.01	0.62 ± 0.003	76.47937	70.12612	42.828	43.478
MesoH_ECS	0.15 ± 0.04	0.14 ± 0.014	56.73466	53.32846	8.510	7.466
μH75_ECS	7.14 ± 0.21	4.52 ± 0.038	− 2.58674	− 2.75695	−18.469	−12.461
μH95_ECS	7.79 ± 0.20	4.60 ± 0.040	− 2.41762	− 2.71579	−18.833	−12.493
μH100_ECS	7.44 ± 0.19	4.54 ± 0.040	− 0.83992	− 0.67257	− 6.249	− 3.053
μH105_ECS	7.31 ± 0.19	4.58 ± 0.040	− 1.62481	− 1.76077	−11.877	− 8.064
Hmax75_ECS	4.42 ± 0.12	4.19 ± 0.037	2.79939	2.80459	12.373	11.751
Hmax95_ECS	4.55 ± 0.11	4.60 ± 0.035	2.49332	2.53503	11.345	11.661
Hmax100_ECS	4.66 ± 0.11	4.27 ± 0.035	2.62547	2.52518	12.235	10.783
Hmax105_ECS	4.41 ± 0.11	4.22 ± 0.036	1.74741	1.96901	7.706	8.309
Constant			−84.83749	−73.69888		

occurs because most proteins imported into the mitochondria need a targeting sequence and are located in the matrix, inner membrane or intermembrane space, and the 47 parameters are mainly related to sequences that possess mitochondrial-targeting sequences. Increasing the training group to include chloroplast proteins (considering that chloroplast proteins can have properties that are similar to mitochondrial ones [6]) did not improve the success rate of the algorithm (data not shown). In conclusion, DFM seems to accurately predict mitochondrial proteins that use a transit peptide to be imported into mitochondria.

According to the weight of each parameter in each LDF (Table 1), the major contribution (22−99-fold) was given by Coef20, H17_KD, MesoH_GvH1 and H17_ECS, which suggests that the amino-acid-residue composition and the structural constraints were extremely important. The parameters μH75_ECS, μH95_ECS, Hmax75_ECS, Hmax95_ECS and Hmax100_ECS contribute in a lesser degree (10−19-fold), which suggested that amphiphilicity is less important. Those that contributed the least (0.02−1.3-fold) were ChDiff, KR, DE, CoefTot, CleavSite, MesoH_KD, μH105_KD and

Table 2. Classification by means of DFM of the SwissProt database sequences. The total number of sequences in each group is indicated in parentheses.

Prediction	Locations of proteins from			
	SwissProt database release 31		SwissProt update (23 October 95)	
	mitochondrial	non-mitochondrial	mitochondrial	non-mitochondrial
	%			
Non-mitochondrial	19.60 (119)	88.67 (11023)	24.53 (13)	88.90 (1405)
Mitochondrial	80.40 (488)	11.33 (1409)	75.47 (40)	11.10 (206)

Hmaxδ_GvH1. The contribution of each parameter distinguished between both groups of proteins by calculating the ratio between mitochondrial and non-mitochondrial contributions. This showed that most of the parameters contributed equally in both cases (ratios of 0.8−1.4). DE, CleavSite, μH105_KD, Hmax105_GvH1 and μH100_ECS parameters, which, except for μH100_ECS, are included among those that contribute the least to the final $P(O/g)$, contributed more for the mitochondrial group (ratios of 1.6−21.5). No parameter was found to contribute more for the non-mitochondrial group. This finding could indicate that for intermediate cases, the particular amino-acid-residue composition, the existence of a cleavage site, and random amphiphilicity are the deciding characteristics.

Mitochondrial-targeting-sequence localization. The possibility to study mitochondrial proteins with some of the described parameters has been published previously [16]. The work has been continued here to develop an algorithm that facilitates the determination of the transit-peptide localization. The sequences used to deduce these empirical rules were chosen as sequences that have in their description, statements that they are mitochondrial precursor proteins coded in the nucleus, their product located in the mitochondria, and use a transit peptide to be imported by the mitochondria. Similar proteins were removed, and only sequences whose final localization is the matrix or inner membrane were chosen. 119 sequences remained, which were divided into two groups, the training data with two thirds of the sequences (80) and the testing data with the rest (39). The testing group was incremented with 131 sequences located in non-mitochondrial cell compartments.

The algorithm empirically developed [MTSL (mitochondrial-targeting-sequence localization)] was divided in four main steps. The first was based on the finding that mitochondrial proteins should not present physical constraints for import, independently of the presence or absence of a putative mitochondrial-targeting sequence. This can be assessed with H17 and MesoH parameters and the hydrophobicity scales GES and KD. The limiting function with GES was defined as an ellipse [12] while that with KD was a line [16]. Sequences with values over the limiting function with at least one scale, were eliminated because this indicates that they present physical constraints to mitochondrial import and cannot be imported, even if they present a targeting-sequence-like structure.

The rationale of the second step is that once a protein has no physical constraint, it should present an N-terminal mitochondrial-targeting sequence that has the following characteristics [5, 16]: they are enriched in positively charged residues (KR) and devoid of acidic ones (DE); and the minimum targeting-sequence length (ZoneTo) is 12 amino acid residues (the minimum length to provide a reliable μHδ). Thus, only sequences with ZoneTo >12 and KR−DE <2 were retained.

The third step is based in the amphiphilic character of the targeting sequences. The ECS was used to compute μHδ for $\delta = 95°$ and $\delta = 100°$ and their corresponding Hmaxδ. Different sets of limiting values based in the mean values of these parameters (Table 1) were tested, but better results were obtained with μH$\delta > 7.3$ and Hmax$\delta > 4.4$, as previously used (see above; [4, 16]). The minimum μHδ and Hmaxδ necessary to provide enough amphiphilicity for a targeting sequence were 5.0 and 3.9, respectively, where the last parameter was determined empirically from the testing group. The first value was changed according to the mean values of Table 1, but final predictions always had less successful rates.

In the last step, to enable the classification, the following were considered: the presence of a cleavage site (CleavSite $\neq 0$); the net charge over the complete amino acid sequence (ChDiff >0); and the amino acid composition, evaluated considering that a mitochondrial-targeting sequence should fulfil that $0.59\times$Coef20 + $0.51\times$CoefTot >0 (where the coefficients are the frequency of agreement of Coef20 and CoefTot, and the real mitochondrial localization).

The sequences were classified in three groups to infer the mitochondrial-targeting sequence: non-mitochondrial, certain mitochondrial and putative mitochondrial. Non-mitochondrial sequences do not fulfil any condition in the first two steps. Sequences that had the minimal μHδ and Hmaxδ but did not fulfil any parameter computed in the last step were also included. Certain mitochondrial sequences had μH$\delta > 7.3$ and Hmax$\delta > 4.4$ (values that qualify a region as sufficiently amphiphilic) and fulfilled some of the characteristics in the fourth step. All other proteins were classified as putative mitochondrial. Only if a protein was qualified as putative or certain mitochondrial, was it expected to have a targeting sequence. Hence, when CleavSite $\neq 0$, the targeting sequence may be assigned up to the position of the last cleavage site. Otherwise it could be assigned to the ZoneTo length. In this region, the part that had the highest μH95 or μH100 marked the region responsible for the initial targeting of the cytoplasmic precursor to the mitochondrion [16]. Several targeting sequences cannot be predicted exactly, because the rules for the cleavage site are not 100% reliable [7], and the most ambiguous rule is not considered in this work, although the parameter ZoneTo is in overall agreement with experimental results.

The training group used to optimize the results of MTSL gave the results, after a jack-knife bias correction, shown in Table 3. The validation was carried out with a testing group composed of 39 mitochondrial sequences and 131 non-mitochondrial sequences. The algorithm seems very reliable, since mito-

Table 3. Prediction of mitochondrial-targeting sequences with MTSL. Values have been bias corrected with the jack-knife method. The total number of sequences in each group are indicated in parentheses. The values for Mitochondrial prediction is the sum of Putative and Certain values.

Prediction	Accuracy of prediction for			
	mitochondrial proteins		non-mitochondrial proteins	
	training group	testing group	including chloroplast proteins	excluding chloroplast proteins
	%			
Non-mitochondrial	8.77 (7)	5.31 (2)	76.28 (101)	84.25 (87)
Mitochondrial	91.23	94.69	23.72	15.75
Putative	38.61 (31)	45.92 (18)	16.48 (21)	11.15 (11)
Certain	52.62 (42)	48.77 (19)	7.24 (9)	4.60 (5)

Table 4. Application of DFM to the sequences used for MTSL. The percentages have been bias corrected with the jack-knife method. The number of sequences in group are in parentheses by the percentage.

Prediction	Accuracy of prediction for			
	mitochondrial proteins		non-mitochondrial proteins	
	training group	testing group	including chloroplast proteins	excluding chloroplast proteins
	%			
Non-mitochondrial	5.24 (4)	2.46 (1)	76.79 (102)	87.92 (91)
Mitochondrial	94.76 (76)	97.54 (38)	23.21 (29)	12.08 (12)

chondrial training and testing groups provided success rates of 91.23% and 94.69%, respectively, which were not statistically different ($P < 0.79$). However, the non-mitochondrial testing group (76.28%) was poorer ($P < 0.34$). If the 28 chloroplast proteins were excluded from this group, the success rate (84.25%) was not statistically different from that obtained with the training group ($P < 0.62$) nor the mitochondrial testing group ($P < 0.46$). This result strengthens the idea that mitochondrial and chloroplastic presequences possess similar structures [6]. In summary, the MTSL algorithm seems to accurately predict the targeting sequence of proteins, provided that chloroplast ones are excluded.

Reliability of the computational methods. DFM needed to be validated with a group of proteins not used to calculate the coefficients. The sequences included in the 23 October 95 update file of the SwissProt database were used as the testing group (Table 2). These results were not statistically different ($P < 0.62$) to those obtained with the complete database. The MTSL algorithm can be interpreted as a simple way to infer whether a protein is mitochondrial or not, similar to the discriminant analysis, which groups the putative and certain classes. Hence, the application of DFM to the proteins of Table 3 gives the results shown in Table 4. The extent of success is always greater for DFM than MTSL. A Student's t test showed that DFM is slightly more accurate than MTSL ($0.025 < P < 0.01$). This is not surprising since MTSL has not been designed to predict mitochondrial proteins but to localize their targeting sequence. Once again, if chloroplast proteins were included, the success rates were lower ($P < 0.20$). All these results support the proposal that DFM can be used with any amino acid sequence to predict whether a protein is localized in the mitochondria or not, and that mitochondrial-targeting sequences inferred with MTSL are reliable.

Predictions of unknown sequences. The application of DFM to sequences from *Saccharomyces cerevisiae* chromosome III [29] is given as an example of the utilization of the algorithms to distinguish between mitochondrial and non-mitochondrial proteins. A prediction on these sequences has been made previously [30]. When both predictions agree that a protein is mitochondrial (for example NFS1, MSH3, YCL68C, YCL34W, YCL33C), it can be strongly suggested that these proteins are really mitochondrial. One of them, MSH3, has been described as having mitochondrial and cytoplasmic isoforms of this protein already in existence [31]. TUP1 is mistakenly predicted as mitochondrial in [30] but DFM predicts that this protein does not belong to the mitochondrial group, which is in agreement with its nuclear localization [32]. Moreover, RM32 is known to be mitochondrial, as predicted by DFM. The differences observed among the predictions of both algorithms can be explained because they are very different algorithms: Coef20 and the existence of a cleavage site are the only criteria that were utilized in the other method [17] for mitochondrial proteins, but are only 2 of the 47 parameters that DFM computes.

By extension of the prediction to all the unidentified *S. cerevisiae* ORF found in the SwissProt database, there seemed to be several clusters of mitochondrial proteins distributed along the

chromosomes. Many mitochondrial proteins are divergently localized, which suggests the possibility that they can share several transcription signals. The rest are convergent or tandem genes. The clustered sequences were selected because there are many genes with related functions that stay together in the genome. Moreover, on chromosome I, YAL034 and YAL035 correspond to the *FUN19* and *FUN12* genes, which encode hypothetical transcription factors. However, most of the predicted ORF are randomly distributed through the chromosomes. This prediction needs to be supported by further work. The clustered genes cited above could be good candidates to start the study. More clusters can be expected because ORF of known function are named differently in the database and can escape our predictions.

Conclusion. We have found 47 significant parameters that will help to predict, by means of a linear approach and a large set of proteins extracted from the SwissProt database, whether a protein is mitochondrially imported or not, when only the amino acid sequence is known. Some of the 47 parameters were enough to localize the mitochondrial-targeting sequences in their amino acid sequences. The use of chloroplast proteins as an equivalent to mitochondrial proteins, according to similarities in their targeting sequence and their import pathways, does not improve the prediction rate. This suggests that although the transit peptides of both kinds of proteins are similar, several differences among them remain undetermined, such as the interaction of the targeting sequence with the distinct lipid composition of chloroplast and mitochondrial membranes [33] or that chloroplast proteins usually need something in addition to the transit peptide [34]. It can be inferred that the information for intracellular sorting is contained in the primary structure of proteins because sequences without post-transcriptional modifications can be used to predict their cell location. Moreover, all the proposed characteristics for mitochondrial proteins have been proven in this work, by means of a large database of proteins. It can be assumed that the transit peptides should be localized among the 40 N-terminal residues. From the predicted mitochondrial proteins in the yeast genome, it can be suggested that many mitochondrially imported proteins are clustered through the chromosomes. In summary, DFM and MTSL are useful tools to study the cell localization of unknown sequences.

The authors would like to acknowledge Dr. C. Jacq and Dr. F. M. Cánovas, for their support and for allowing the development of part of this work in their laboratories, and E. Sterling and J. Garzon, for their English revision. MGC was supported by a *Plan de Formación del Personal Investigador* postdoctoral fellowship from the Spanish *Ministerio de Educación y Ciencia.*

REFERENCES

1. Pfanner, N., Söllner, T. & Neupert, W. (1991) Mitochondrial import receptors for precursor proteins, *Trends Biochem. Sci. 16*, 63–67.
2. Verner, K. (1992) Early events in yeast mitochondrial protein targeting, *Mol. Microbiol. 6*, 1723–1728.
3. Kiebler, M., Becker, K., Pfanner, N. & Neupert, W. (1993) Mitochondrial protein import: specific recognition and membrane translocation of preproteins, *J. Membr. Biol. 135*, 191–207.
4. von Heijne, G. (1986) Mitochondrial targeting sequences may form amphiphilic helices, *EMBO J. 5*, 1335–1342.
5. Schatz, G. (1987) Signals guiding proteins to their correct locations in mitochondria, *Eur. J. Biochem. 165*, 1–6.
6. von Heijne, G., Steppuhn, J. & Herrmann, R. G. (1989) Domain structure of mitochondrial and chloroplast targeting peptides, *Eur. J. Biochem. 180*, 535–545.
7. Gavel, Y. & von Heijne, G. (1990) Cleavage-site motifs in mitochondrial targeting peptides, *Protein Eng. 4*, 33–37.
8. Pack, Y. K. & Weiner, H. (1990) Import of chemically synthesized signal peptides into rat liver mitochondria, *J. Biol. Chem. 265*, 14298–14307.
9. Bruch, M. D. & Hoyt, D. W. (1992) Conformational analysis of a mitochondrial presequence derived from the F_1-ATPase β-subunit by CD and NMR spectroscopy, *Biochim. Biophys. Acta 1159*, 81–93.
10. Wang, Y. & Weiner, H. (1993) The presequence of rat liver aldehyde dehydrogenase requires the presence of an α-helix at its N-terminal region which is stabilized by the helix at its C-terminal, *J. Biol. Chem. 268*, 4759–4765.
11. Jaussi, R. (1995) Homologous nuclear-encoded mitochondrial and cytosolic isoproteins. A review of structure, biosynthesis and genes, *Eur. J. Biochem. 228*, 551–561.
12. Claros, M. G., Perea, J., Shu, Y., Samatey, F. A., Popot, J. L. & Jacq, C. (1995) Limitations of the *in vivo* import of hydrophobic proteins into yeast mitochondria. The case of a cytoplasmically synthesized apocytochrome *b*, *Eur. J. Biochem. 228*, 762–771.
13. Claros, M. G., Perea, J. & Jacq, C. (1996) Allotopic expression of a yeast mitochondrial maturase to study mitochondrial import of hydrophobic proteins, *Methods Enzymol. 264*, 389–403.
14. Baker, A. & Schatz, G. (1987) Sequences from a prokaryotic genome or the mouse dihydrofolate reductase gene can restore the import of a truncated precursor into yeast mitochondria, *Proc. Natl Acad. Sci. USA 84*, 3177–3121.
15. Hurt, E. C. & Schatz, G. (1987) A cytosolic protein contains a cryptic mitochondrial targeting signal, *Nature 325*, 499–503.
16. Claros, M. G. (1995) MitoProt, a Macintosh application for studying mitochondrial proteins, *Comput. Appl. Biosci. 11*, 441–447.
17. Nakai, K. & Kanehisa, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells, *Genomics 14*, 897–911.
18. Bairoch, A. Boeckmann, B. (1994) The SWISS-PROT protein sequence data bank: current status, *Nucleic Acids Res. 22*, 3578–3580.
19. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986) Identifying nonpolar transbilayer helices into amino acid sequences of membrane proteins, *Annu. Rev. Biophys. Biophys. Chem. 15*, 321–353.
20. von Heijne, G. (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, *J. Mol. Biol. 225*, 487–494.
21. Kyte, J. & Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol. 157*, 105–132.
22. Eisenberg, D. (1984) Three-dimensional structure of membrane and surface proteins, *Annu. Rev. Biochem. 53*, 595–623.
23. Kendall, M. G. & Stuart, A. (1968) *The advanced theory of statistics: design and analysis, and time-series*, Charles Griffin Company Limited, London.
24. Tarroux, P., Vincens, P. & Rabilloud, T. (1987) HERMES: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part V: data analysis, *Electrophoresis 8*, 187–199.
25. Efron, B. & Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, *Stat. Sci. 1*, 54–77.
26. DeGroot, M. H. (1986) *Probability and statistics*, Addison-Wesley Publising Company, Inc., Reading.
27. Swanson, S. T. & Roise, D. (1992) Binding of a mitochondrial presequence to natural and artificial membranes. Role of surface potential, *Biochemistry 31*, 5746–5751.
28. Klein, P., Kanehisha, M. & DeLisi, C. (1985) The detection and classification of membrane-spanning proteins, *Biochim. Biophys. Acta 815*, 468–476.
29. Oliver, S. G., et al. (1992) The complete DNA sequence of yeast chromosome III, *Nature 357*, 38–46.
30. Slonimski, P. P. & Brouillet, S. (1993) A data-base of chromosome III of *Saccharomyces cerevisiae*, *Yeast 9*, 941–1029.

31. New, L., Liu, K. & Crouse, G. F. (1993) The yeast gene *MSH3* helps define a new class of eukaryotic *MutS* homologous, *Mol. & Gen. Genet 329*, 97–108.
32. Williams, F. E., Varanasi, U. & Trumbly, R. J. (1991) The CYC8 and TUP1 proteins involved in glucose repression in *Saccharomyces cerevisiae* are associated in a protein complex, *Mol. Cell. Biol. 11*, 3307–3316.
33. de Kruijff, B. (1994) Anionic phospholipids and protein translocation, *FEBS Lett. 346*, 78–82.
34. de Castro Silva Filho, M., Chaumont, F., Leterme, S. & Boutry, M. (1996) Mitochondrial and chloroplast targeting sequences in tandem modify protein import specificity in plant organelles, *Plant Mol. Biol. 30*, 769–780.

Supplementary material. **Computational method to predict mitochondrially imported proteins and their targeting sequences.**
Table S1. Mitochondrial-targeting-sequence prediction of 18 mitochondrial proteins considered as 'putative' or 'certain'.
Table S2. Results of the prediction of chromosome III sequences made with two algorithms.
Table S3. Mitochondrial ORF clustered along the yeast chromosomes. This information is available, on request, from the Editorial Office. Nine pages are available.