Diversity in Degrees of Freedom of Mitochondrial Transit Peptides

Christine Staiger,†[1,2] Alexander Hinneburg,*[2] and Ralf Bernd Klösgen†*

*Faculty of Science III, Institute of Computer Science, Martin-Luther-University Halle-Wittenberg, Halle/Saale, Germany; and † Faculty of Science I, Institute of Biology—Plant Physiology, Martin Luther University Halle-Wittenberg, Germany

Most mitochondrial proteins are synthesized in the cytosol of eukaryotic cells as precursor proteins carrying N-terminal extensions called transit peptides or presequences, which mediate their specific transport into mitochondria. However, plant cells possess a second potential target organelle for such transit peptides, the chloroplast. It can therefore be assumed that mitochondrial transit peptides in plants are exposed to an increased demand of specificity, which in turn leads to reduced degrees of freedom in these transit peptides compared with those of nonplant organisms. Our study investigates this hypothesis using fractal dimension. Statistical analysis of sequence data shows that the fractal dimension of mitochondrial transit peptides in plants is indeed significantly lower than that from nonplant organisms.

Introduction

Mitochondria are of endosymbiotic origin, that is, they are derived from the engulfment of a bacterium into a hitherto unknown host cell, an event that finally resulted in the development of eukaryotic cells. In the course of evolution, most of the mitochondrial genes were transferred to the nucleus. As a consequence, most mitochondrial proteins are synthesized in the cytosol of the cell as precursor polypeptides carrying cleavable aminoterminal extensions, named presequences or transit peptides, that mediate transport of the protein "back" into the organelle.

Comparison of such mitochondrial transit peptides has demonstrated that they 1) can be quite variable in size; 2) contain many positively charged, hydrophobic, and hydroxylated amino acid residues; and 3) have a high tendency to form an amphipathic α-helix (von Heijne et al. 1989; Pfanner and Geissler 2001). However, because most of these comparisons are based on mitochondrial transit peptides from fungi and mammals (see, e.g., von Heijne et al. 1989), these conclusions might well be biased and must not necessarily hold true for all species. This is particularly obvious for plants because plant cells harbor an additional class of organelles of endosymbiotic origin, notably chloroplasts (or generally speaking, plastids). These organelles originate from a second endosymbiotic event in which a cyanobacterium was engulfed by a eukaryotic host cell possessing already mitochondria. It can be assumed that the evolutionary establishment of chloroplasts had an effect also on the selection pressure operating on mitochondrial transit peptides because these transport signals were suddenly exposed to the situation that a second potential target organelle was present within the same cell. The situation was further complicated by the fact that also most chloroplast genes were phylogenetically transferred to the nucleus. Again, cleavable transit peptides for the transport of the corresponding proteins back into the organelles were developed, which show remarkable similarity to mitochondrial transport signals in terms of N-terminal position and amino acid composition. Considering this scenario, one could predict that mitochondrial transit peptides of plant cells must have adapted to this new situation by developing a higher degree of specialization in order to prevent permanent transport into the wrong organelle. Supporting evidence for this assumption comes from the observation that several nuclear encoded proteins show dual targeting into both mitochondria and chloroplasts because they carry transit peptides with ambiguous organelle specificity (for a recent review, see Carrie et al. 2009). The number of proteins identified with such targeting properties has significantly increased in the past years suggesting that this is a much more common phenomenon than originally anticipated. Still, it can be assumed that mistargeting is only to some degree tolerable for the cell and will probably disturb its integrity and the division of labor between the organelles if it exceeds a certain level.

These considerations led us to the following working hypothesis: although mitochondrial transit peptides in general are characterized by high degrees of freedom (df) in terms of amino acid sequence and composition, plant mitochondrial transit peptides should be significantly less variable in this respect. In order to examine this hypothesis experimentally, mitochondrial transit peptides from one model species each of plants (*Arabidopsis thaliana*), mammalia (*Mus musculus*), and fungi (*Saccharomyces cerevisiae*) were compared by a bioinformatic approach. We used for this purpose fractal dimension, which is a measure of complexity of sequence information within a group of sequences.

Materials and Methods

Estimating the df of Protein Sequences

The estimation of the df of a set of protein sequences is a nontrivial task. We interpret the df as the number of independent dimensions, which are necessary to span the data space. In our case, the data space is embedded into the space of all sequences of length m over the alphabet of the 20 proteinogenic amino acids $\Sigma = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$. The m dimensions correspond to the sequence positions $1,\ldots,m$, which can take either of the different amino acids as values.

In our case, the sequences to be analyzed are transit peptides, which are represented by the 50 N-terminal

[1] Present address: BioSystems Research Group, School of Computing, University of Leeds, Leeds, United Kingdom.

[2] Both authors contributed equally to this work.

Key words: mitochondria, transit peptide, degrees of freedom, fractal dimension, target organelle, organelle specificity.

E-mail: klosgen@pflanzenphys.uni-halle.de.

Mol. Biol. Evol. 26(8):1773–1780. 2009
doi:10.1093/molbev/msp087
Advance Access publication April 22, 2009

amino acids of the mitochondrial precursor proteins. In order to constitute a functional transit peptide, not all 50 positions can be filled arbitrarily with amino acids. Instead, some restrictions do apply which are, however, not yet explicitly known. Thus, not each of the positions is counting as an independent dimension and the true dimensionality, which also accounts for the unknown restrictions, is probably less than 50. We avoid the explicit estimation of those restrictions from sequence data, which demands large data sets. Instead, the true dimensionality is directly estimated from sequence similarities.

We draw on the concept of correlation dimension, which is one of several definitions of fractal dimension. The concept of fractal dimension (Mandelbrot 1977; Falconer 1990) has different theoretical definitions among which the Hausdorff dimension is a prominent one. The general idea is to cover the data space by a number of nonempty balls. For practical implementations, the definition of correlation dimension, which we are going to describe in this section, is more useful.

Correlation Dimension

Let $X = \{x_1, x_2, \ldots\}$ be a set of objects and d_{ij} be the distance between x_i and x_j. The correlation integral $C(r)$ for a given radius r is

$$C(r) = \limsup_{N \to \infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(r - d_{ij}). \tag{1}$$

The specific use of the Heaviside function $\Theta(b) = \begin{cases} 0 & \text{, if } b \leqslant 0 \\ 1 & \text{, else} \end{cases}$ indicates whether two objects have a distance smaller than r. So, the inner sum over j counts how many objects are in a ball of radius r centered at x_i. The term after the limit denotes the fraction of pairs of vectors with an index smaller N and which are closer than r. The limit is derived by assuming an infinite set of objects with respective distances. The correlation dimension d is the rate at which the logarithm of the correlation integral decreases when r is shrinking.

$$d = \lim_{r \to 0} \frac{\log C(r)}{\log r}. \tag{2}$$

In practice, only finite data sets are available. A popular heuristic method to estimate correlation dimension from finite data is by Grassberger and Procaccia (1983). When dealing with finite data, the original definition of the correlation integral equation (1) cannot be used due to the limit $N \to \infty$. Note that the equation can be seen as an average (outer sum with an factor of $1/N$) of N fractions each giving the fraction of objects contained in a ball (inner sum with an factor of $1/N$). A standard heuristic explained in Sprott (2003), which is used to get more robust estimates of the fraction of data contained in a ball, is to leave out the center point that induces the ball. This could lead to empty balls for small radii. The empty balls must be filtered out before the computation of the correlation integral in order to avoid distortion of the estimate of the correlation dimension.

We denote by $\hat{p}_i$ the estimate of the fraction of data contained in the ith ball of radius r:

$$\hat{p}_i(r) = \frac{1}{N-1} \sum_{j=1, i \neq j}^{N} \Theta(r - d_{ij}). \tag{3}$$

In order to avoid empty balls in the following formulas, we denote by $\hat{N}(r)$ the number of nonempty balls, which is basically the number of $\hat{p}_i$s that are larger than zero. Note that $\hat{N}(r) \leqslant N$. Additionally, let $\hat{I}(r) \subseteq \{1, \ldots, N\}$ the index set of those nonzero $\hat{p}_i(r)$s.

The Grassberger–Procaccia algorithm (Grassberger and Procaccia 1983) estimates the correlation dimension by computing an estimate of the correlation integral for several representative values $r_1, r_2, \ldots$ for r. In order to estimate the correlation integral for a fixed r, the algorithm computes the average of the $\hat{N}(r)$ nonzero fraction estimates $\hat{p}_i(r)$.

$$\hat{C}(r) = \frac{1}{\hat{N}(r)} \sum_{i \in \hat{I}(r)}^{\hat{N}(r)} \hat{p}_i(r). \tag{4}$$

Then, $\log \hat{C}(r)$ is plotted versus $\log r$ in the so-called log–log plot and a line is fitted to the points of the linear part of that curve. Correlation dimension is estimated as the slope of the line fitted to the linear part of the curve in the log–log plot.

Simple Examples

We illustrate the idea behind correlation dimension by two simple examples. Assume the given data set is a finite sample of points, which are uniformly distributed in the two-dimensional plane (fig. 1*A*). The number of points in a ball of radius r around a particular point x_i is approximately proportional to the area covered by the ball, which is πr^2. Thus, the correlation integral grows nearly quadratically for medium values of r. Figure 1*C* shows the corresponding log–log plot. The slope is estimated by fitting a line to the marked points. In the log–log plot, the linear part of the lower curve has a slope close to two. This corresponds to the fact that the original data are uniformly distributed in the two-dimensional plane. The tail for small values for r is to be ignored because in this area, the correlation integral depends on balls including only few points. The part for very large radii is likewise not informative because in this case, the balls include all points and, therefore, the correlation integral cannot grow further.

In the second example, points are sampled from a line, which is arbitrarily embedded within the two-dimensional space (fig. 1*B*). The number of points in a ball of radius r around a particular point x_i is approximately proportional to the length of the line covered by the ball. Thus, the correlation integral grows almost linearly. In the corresponding log–log plot, the linear part of the upper curve has therefore a slope close to one (fig. 1*C*).

Application of Fractal Dimensionality to Sequences

A common method to compare biological sequences is to align two sequences and compute a similarity score

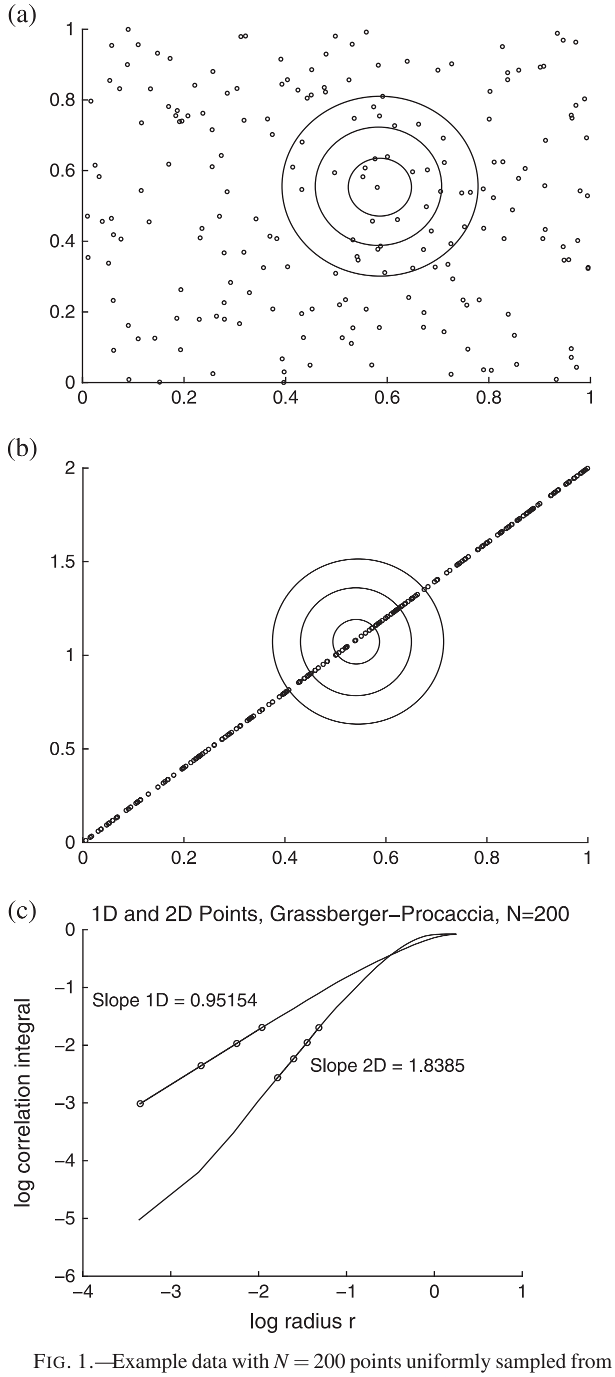

FIG. 1.—Example data with $N = 200$ points uniformly sampled from a plane (*a*) and from a line (*b*). Radius versus correlation integral for both data sets (*c*) and the slopes of the fitted lines, which estimate correlation dimensions. The lines are fitted to the marked points.

from such alignments. Therefore, we have first computed a similarity score based on alignments for each pair of transit peptides sequences. Then, the correlation dimension was calculated from these similarity scores.

Alignments help detect biologically relevant similarities between protein sequences. The basic idea behind an alignment between two sequences A and B is to find a transformation from A to B with maximal similarity score in terms of simple edit operations like insert, delete, match, and mismatch. In order to compute an alignment with maximal similarity score, parameters are needed to specify the scores of basic transformation operations. As we are comparing whole transit peptide sequences, we perform global alignments instead of local ones. For the computation of pairwise alignments, we have applied the standard algorithm ClustalW (Thompson et al. 1994).

In our work, we chose the Blosum$_{62}$ matrix as score matrix for matches and mismatches. Furthermore, gaps are penalized with negative scores for gap opening and gap extension. We used a standard parameter combination for gap opening and gap extension, namely -10 and -0.1, respectively, which represents the default setting in ClustalW.

The direct output of such an alignment is the maximal similarity score of the whole transformation. However, those direct outputs are not comparable for different pairs of sequences. As the computation of the correlation integral averages over the number of objects within balls of the same radii but different centers, such a comparison of similarity scores is implicitly assumed. Pairwise alignment in ClustalW already does such a normalization (Thompson et al. 1994), namely by dividing the number of identical sequence positions in the alignment by the number of matched residues 0 and multiplying by 100. This defines a similarity functions that ranges from 0 to 100.

Correlation dimension is defined in terms of distances instead of similarities. When distances are small, the corresponding similarities are large. Therefore, the definition of correlation integral needs to be adapted to handle similarities. We adapt $\hat{p}_i(r)$ that was previously defined in equation (3) by flipping the difference in the argument of the Heaviside function:

$$\hat{p}_i(r) = \frac{1}{N-1} \sum_{j=1, i \neq j}^{N} \Theta(s_{ij} - r). \tag{5}$$

All other equations remain unchainged.

Data

The data sets used were retrieved from the SwissProt/UNIProt database, release 14.1 (http://www.uniprot.org). All entries from mouse (*M. musculus*), yeast (*S. cerevisiae*), and Arabidopsis (*A. thaliana*) that have a location attribute containing "mitochondrion" and a topic attribute containing "transit peptide" were collected. Note that the data also include proteins that are only predicted to be mitochondrial proteins. However, SwissProt/UNIProt is quite conservative with those annotations. We obtained 319 entries for yeast, 427 for mouse, and 224 for Arabidopsis. Because the lengths of the transit peptides were often not known, the 50 N-terminal amino acids of each protein sequence were taken as putative transit peptides.

For each data set of transit peptides, we computed all normalized pairwise alignment scores by ClustalW (version 1.7) using the slow and more accurate alignment method. Note that the slow alignment methods implemented in ClustalW are essentially the basic methods known as Needleman–Wunsch alignments. As we did not use multiple alignments but pairwise alignments only, ClustalW is sufficient. New alignment tools like T-COFFEE (Notredame et al. 2000) or MUSCLE (Edgar 2004) offer faster approximations of pairwise alignments or more accurate multiple alignments. Both features are not needed in this project.

Thus, in total, a quadratic matrix with normalized pairwise similarities was derived for each set of transit peptides.

Results

In the first experiment, the correlation dimension of each of the three data sets of transit peptides is calculated by the Grassberger–Procaccia algorithm. In order to avoid any bias from the different sizes of the data sets, correlation dimension is not calculated on the full data sets but on random samples of identical size.

We use a more sophisticated sampling method, namely bootstrap sampling (Efron and Tibshirani 1998). The general idea of bootstrap is to construct a bootstrap sample from a given original data set of size *N* by randomly sampling *N* objects with replacement. Because sampling with replacement may choose some data objects more than once, a bootstrap sample may include duplicates. Assuming the original data does not include any duplicates, choosing the objects uniformly with replacement puts on average about 63% unique objects into a bootstrap sample, whereas the rest are duplicates. The bootstrap sampling method allows to build random samples as large as the original data set.

To construct comparable bootstrap samples for each of the three different data sets, the two larger ones (*S. cerevisiae* and *M. musculus*) need to be downsampled to the size of the *A. thaliana* data set, which is $N = 224$. In order to construct a bootstrap sample of size *N* from a data set with size larger *N*, the first step is to draw *N* objects from the original data set without replacement. The actual bootstrap sample is generated in a second step by sampling *N* objects with replacement from the objects drawn in the first step. Both steps are repeated to generate the next bootstrap sample. Using such a two-step procedure instead of directly sampling *N* objects with replacement from the original data set keeps the percentage of unique objects in the bootstrap samples at about 63% of the sample size *N* across all three data sets.

For each estimation of the correlation dimension of a data set, a random bootstrap sample is computed from the original data as described above. The correlation integral is computed for several radii and the logarithm of the similarity radius is plotted versus the logarithm of the correlation integral. A line is fitted to the linear part of that curve and the slope serves as an estimate of correlation dimension.

Figure 2 shows for each of the three data sets the results derived from only five random bootstrap samples. The absolute values of the slopes are shown in the insets of the figures. Because visual comparison across the figures is difficult, representative log–log plots of each data set are combined in figure 3. The values of the slopes calculated for each sample suggest that the transit peptides of *A. thaliana* have a lower correlation dimension than those of *S. cerevisiae* and *M. musculus*.

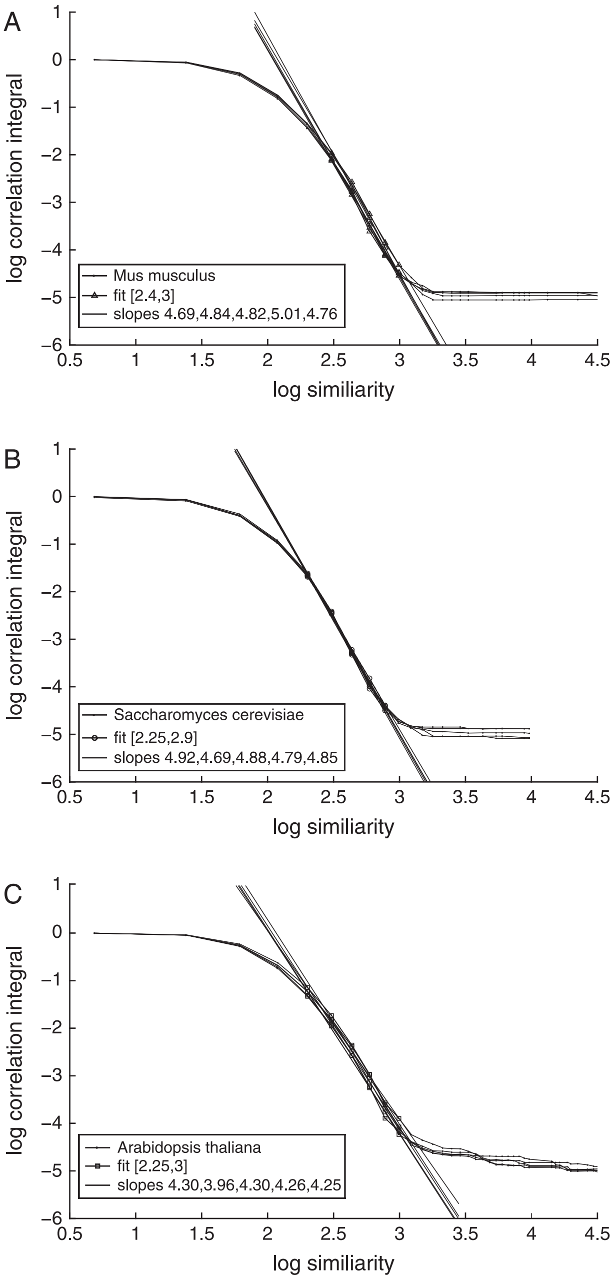

FIG. 2.—Log–log plots and calculated values of correlation dimension (slope) for all data sets, (*A*) *Mus musculus*; (*B*) *Saccharomyces cerevisiae*; and (*C*) *Arabidopsis thaliana*), using ClustalW with 10, 0.1 as gap opening and gap extension parameters, respectively. In each case, five examples are shown. The calculated slopes of the fitted lines in those examples are shown in the insets.

In order to substantiate the results, the calculation was repeated with 1,000 random bootstrap samples each of the transit peptides of *S. cerevisiae*, *M. musculus*, and *A. thaliana*. The means of the calculated correlation dimensions including error bars showing the standard deviations (SDs) derived from the results of the 1,000 random bootstrap samples are depicted in figure 4. Again, it becomes obvious that the mitochondrial transit peptides of *A. thaliana* have a lower correlation dimension than those of *S. cerevisiae* and *M. musculus*.

In order to examine to what extent the results of the calculated correlation dimensions are sensitive to the

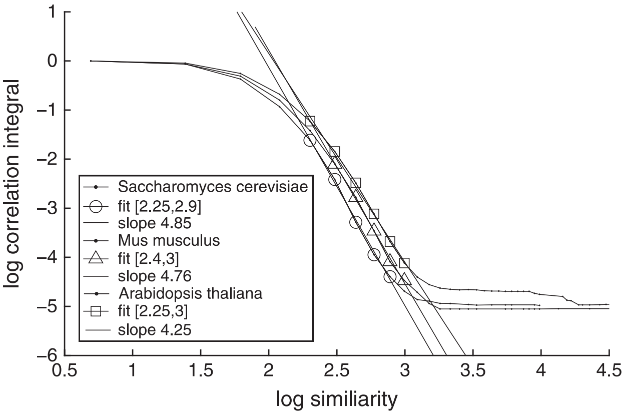

FIG. 3.—Comparison of log–log plots and calculated values of correlation dimension (slope) for all data sets. For further details, see the legend to figure 2.

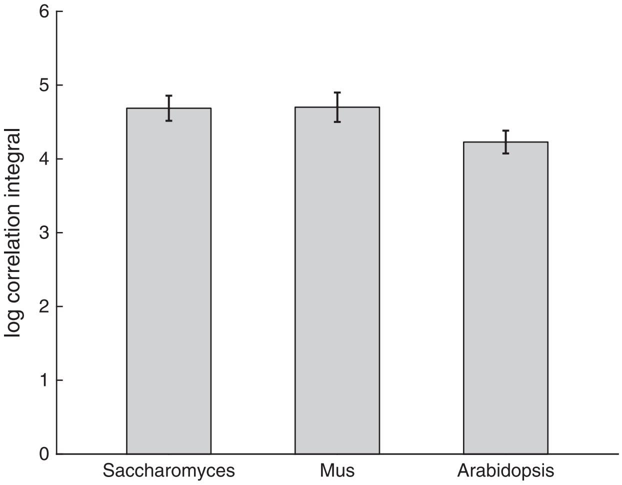

FIG. 4.—Means of the calculated correlation dimensions of *Mus musculus* (Mus), *Saccharomyces cerevisiae* (Saccharomyces), and *Arabidopsis thaliana* (Arabidopsis) taken from 1,000 random bootstrap samples each ($N = 224$). The error bars show the respective SD.

size of the bootstrap samples used, we varied the sample size $N \in \{75, 100, 125, 150, 200, 224\}$. The maximal sample size is determined by the smallest data set, which in our case is that of *A. thaliana*. We generated for each data set and sample size 1,000 bootstrap samples and estimated the correlation dimension for each bootstrap sample. Subsequently, we derived the mean and SD from the 1,000 bootstrap samples generated for each particular sample size. Figure 5*A* shows that with growing sample size, the calculated correlation dimension of the transit peptides of *A. thaliana*, on the one hand, and those of *S. cerevisiae* and *M. musculus*, on the other hand, drift apart.

The visual impression that the transit peptides of *A. thaliana* have a lower correlation dimension than those of *S. cerevisiae* and *M. musculus* is verified by *t*-tests. In general, a statistical test like the *t*-test consists of a null hypothesis, which states the opposite of the observation. Loosely spoken, it plays the role of the “devil's advocate.” In our case, the null hypotheses are that the means of the correlation dimension of the transit peptides of *S. cerevisiae* and *A. thaliana* as well as those of *M. musculus* and *A. thaliana* are equal. Both null hypotheses can safely be rejected considering that the respective *p* values become numerically 0, which is obviously lower than any reasonable standard significance level. This demonstrates the significance of the observation that the correlation dimension of the transit peptides of *A. thaliana* is smaller than those of *S. cerevisiae* and *M. musculus*. As a kind of control, the null hypothesis that the means of the correlation dimension of the transit peptides of *S. cerevisiae* and *M. musculus* are equal is analogously tested. Remarkably, the resulting *p* value is 0.1075 in this instance, which does not even allow to reject the null hypothesis at a significance level of only 10%. Thus, our analysis based on correlation dimension does not reveal significant differences between *S. cerevisiae* and *M. musculus*.

Effect of Data Set Variation

In order to examine if the results described so far have been biased by certain parameters of the data examined, the analysis was repeated with modified data sets. The first modification concerns the size of the selected transit peptides. In the original analysis, we have taken the 50 N-terminal amino acid residues of each protein as the mitochondrial targeting signal because the exact size of the transit peptides was only in few cases experimentally determined. Several transit peptides are, however, shorter than 50 residues, and it must therefore be assumed that the data sets include also significant amounts of mature protein sequences. This might influence the outcome because mature protein sequences are presumably exposed to completely different selective pressure than transit peptides. In order to reduce the potential effect of such mature sequences, we have performed the analysis also with data sets containing the 40 N-terminal residues of each protein only. Though it will lead to C-terminal truncation of those transit peptides that exceed 40 residues, it will first and foremost reduce the contamination with mature protein sequences. Bootstrap analysis performed with these new data sets yields essentially similar results as described above: the transit peptides of *A. thaliana* have a significantly lower correlation dimension than those of *S. cerevisiae* and *M. musculus* (figure 5*B*). Note that the absolute values of the correlation dimension for transit peptides of length 40 are lower than for those of length 50 because shortening of the assumed transit peptides unevitably decreases the variability in the data, which in turn leads to lower absolute values of the correlation dimension.

A second parameter, which might influence the degree of correlation dimension in a given data set are homologous proteins, which are generally the result of gene duplication and therefore bound to be quite closely related, even in their transit peptide sequences. If the number of homologous proteins within one data set differs significantly from those of the other data sets, it might well have a strong influence on the calculated correlation dimension. In order to take this possibility into account, we have strived to eliminate homologous proteins from all three data sets. For this purpose, we computed alignment scores between

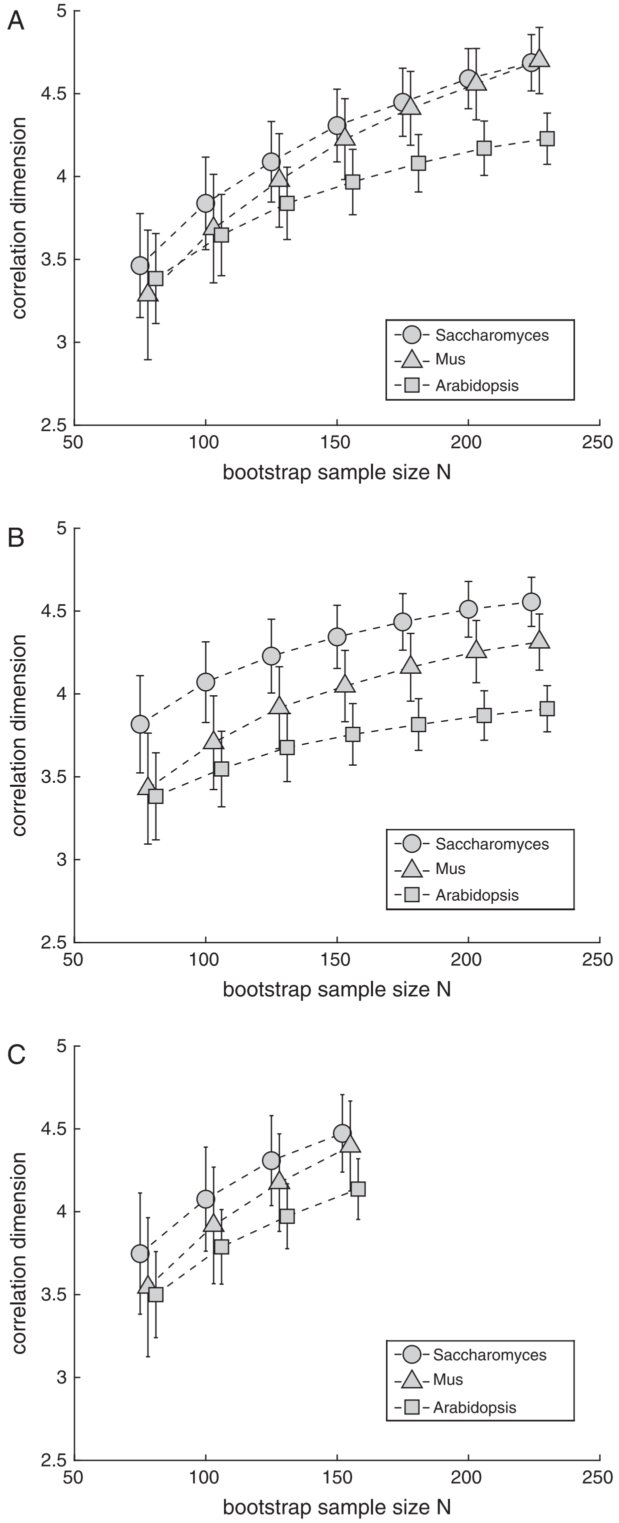

FIG. 5.—Dependency of correlation dimension on the sample size. The data shown are means and SD calculated from 1,000 bootstrap samples each. (*A*) Original data set with the 50 N-terminal amino acid residues taken as transit peptides. (*B*) Data entries as in (*A*) but taking only the 40 N-terminal residues as transit peptides. (*C*) Data set as in (*A*) but devoid of homologous proteins.

Table 1
Normalized Usage Frequencies of Amino Acids in the Total Mitochondrial Protein Sequences

Amino acid	*Arabidopsis thaliana*	*Mus musculus*	*Saccharomyces cerevisiae*
A	0.0764	0.0833	0.0635
C	0.0174	0.0176	0.0107
D	0.0528	0.0462	0.0520
E	0.0670	0.0639	0.0619
F	0.0407	0.0371	0.0424
G	0.0711	0.0721	0.0577
H	0.0209	0.0255	0.0213
I	0.0564	0.0476	0.0644
K	0.0665	0.0577	0.0814
L	0.0947	0.1030	0.0985
M	0.0304	0.0241	0.0219
N	0.0398	0.0316	0.0534
P	0.0421	0.0544	0.0453
Q	0.0289	0.0424	0.0380
R	0.0545	0.0628	0.0505
S	0.0788	0.0677	0.0775
T	0.0504	0.0519	0.0572
V	0.0725	0.0707	0.0602
W	0.0096	0.0131	0.0104
Y	0.0291	0.0273	0.0319

the full protein sequences within each data set and built groups using the single linkage algorithm (Sibson 1973). The groups are built such that no sequence entry from two different groups exceeds the low sequence similarity score of 30 (ClustalW assigns to each pair of sequences a score between 100 [very similar] and 0 [not similar]). Thus, elements of different groups have quite divergent sequences and are considered to be nonhomologous. For the subsequent bootstrap experiments in which *N* groups are picked randomly, only one representant from each group is randomly chosen. Thus, each bootstrap sample contains at most one entry from a given group, which prevents that homologous entries are present when computing the correlation dimension of a bootstrap sample. Using the cutoff value of 30 for sequence similarity, 154 groups are defined for the smallest data set (Arabidopsis), which in turn also limits the maximal sample size for this experiment to 154. The results show that even after elimination of duplicates, the transit peptides of *A. thaliana* have a lower correlation dimension than those of *S. cerevisiae* and *M. musculus* (figure 5*C*). Due to the smaller maximal sample size of 154, the difference is not as pronounced as with the maximal sample size of 224 of the original experiment (figure 5*A*), but it is still statistically significant with numerically zero *p* values. Actually, if identical sample sizes are compared between the different experiments, the differences in variability between the transit peptides of Arabidopsis, mouse, and yeast are comparable in the two experiments (cf. figures 5*A* and *C*). The absolute values of the correlation dimensions in the data sets devoid of homologous proteins are slightly increased compared with the corresponding values in the original experiments though, due to the lack of redundancy in the data sets, which slightly increases the variability and, in turn, the correlation dimension.

Finally, the usage of amino acid residues in total protein sequences (transit peptide plus passenger protein) is analyzed to ensure that the observed effect is not caused by different amino acid preferences in the three organisms. The normalized usage frequencies are computed from pooled sequences, that is, all sequences are concatenated and the normalized frequencies are computed as the number of occurrences of a particular amino acid divided by the total length of the concatenated sequence.

The normalized usage frequencies are shown in table 1. Except for Q and W, which are both quite rare amino acids in proteins and can thus not be responsible for the observed differences in correlation dimension, the normalized usage frequency of amino acid in *A. thaliana* is close to those of *S. cerevisiae* and *M. musculus*. This is in line with the assumption that the overall usage of amino acids is similar in all three organisms and confirms that *A. thaliana* has no general bias in the amino acid usage. Thus, it must be concluded that the significantly lower correlation dimension of mitochondrial transit peptides of *A. thaliana* is a consequence of reduced df in the composition of these protein transport signals.

Discussion

During the past decade, several algorithms were developed to predict the subcellular localization of proteins by examining their N-terminal targeting sequences. Examples are the neural network–based approaches TargetP (Emanuelsson et al. 2000) and Predotar (Small et al. 2004). Both examine the 100 N-terminal amino acids and learn from training examples to discriminate between mitochondrial transit peptides, chloroplast transit peptides, and other transport signals. TargetP predicts a score for each of the first 100 amino acids giving a likelihood whether it represents a transport signal and classifies on that basis the target organelle of the protein. Predotar uses information on net charge, hydrophobicity, and amino acid distribution to predict the target organelle. Other approaches classifying proteins according to their targeting sequence are PSORT and MitoProtII. PSORT (Nakai and Horton 1999) is a rule-based expert system and computes the likelihood that a given protein belongs to a specific target. MitoProt II (Claros and Vincens 1996) can only distinguish between mitochondrial and nonmitochondrial transport signals.

Neither of these analyses has considered the particularities of plant transit peptides. Instead, the training data used by the described algorithms collect mitochondrial transit peptides from plants, animals, and fungi within a single group. Thus, the potential differences between mitochondrial transit peptides of plants and animals or fungi are neglected. This position is supported in Emanuelsson et al. (2000) by citing a cluster analysis (Schneider et al. 1998), which found no species-correlated differences between mitochondrial transit peptides. However, the data basis of that study used only 14 transit peptides from plants among 144 transit peptides in total, which does not allow any statistical conclusions. In contrast, our results strongly suggest that there are species-dependent differences among mitochondrial transit peptides. Thus, the predictions obtained by TargetP and Predotar have to be reconsidered when analyzing sequences from plants. It is furthermore remarkable that both programs are used to annotate the transit peptides in the Uniprot/Swissprot database.

Information theoretic methods have already been used before to analyze biological phenomena. However, although those analyses often try to find commonalities between different sequences or regions of sequences using mutual information, for example, to describe molecular coevolution (Codoñer and Fares 2008), our approach is based on fractal dimension, which detects complexity differences between data sets.

To our knowledge, our study is the first one investigating the hypothesis that mitochondrial transit peptides of plants are more specialized and consequently have less df than those of animals or fungi. This hypothesis is tested by estimating correlation dimension of sets of transit peptides from three example organisms. Our results show that the correlation dimension of transit peptides from *A. thaliana* is significantly lower than that from *M. musculus* and *S. cerevisiae*, in line with the assumption that plant mitochondrial transit peptides are exposed to increased selective pressure concerning organelle specificity. In future work, these analyzes will be extended to further organisms to evaluate the significance of this observation for plants in general.

Acknowledgments

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 648). We thank Ludwig Staiger for helpful comments during the preparation of the manuscript.

Literature Cited

Carrie C, Giraud E, Whelan J. 2009. Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts. FEBS J. 276:1187–1195.

Claros MG, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem. 241:779–786.

Codoñer FM, Fares MA. 2008. Why should we care about molecular coevolution? Evol Bioinform. 4:29–38.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Efron B, Tibshirani RJ. 1998. An introduction to the bootstrap. Boca Raton, New York, and Washington D.C. (US): Chapman & Hall.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. J Mol Biol. 300:1005–1016.

Falconer KJ. 1990. Fractal geometry. Chichester (United Kingdom): Wiley.

Grassberger P, Procaccia I. 1983. Estimation of the Kolmogorov entropy from a chaotic signal. Phys Rev A. 28:2591–2593.

von Heijne G, Stepphuhn J, Herrmann RG. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. Eur J Biochem. 180:535–545.

Mandelbrot BB. 1977. Fractals, form, chance, and dimension. San Francisco (CA): Freeman.

Nakai K, Horton P. 1999. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci. 24:34–36.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

Pfanner N, Geissler A. 2001. Versatility of the mitochondrial protein import machinery. Nat Rev Mol Cell Biol. 2:339–349.

Schneider G, Sjöling S, Wallin E, Wrede P, Glaser E, von Heijne G. 1998. Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. Proteins: Struct Funct Genet. 30:49–60.

Sibson R. 1973. Slink: an optimally efficient algorithm for the single-linkage cluster method. Comput J. 16:30–34.

Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: a tool for rapidly screening proteomes for n-terminal targeting sequences. Proteomics. 4:1581–1590.

Sprott JC. 2003. Chaos and time-series analysis. Oxford (United Kingdom), and New York (US): Oxford University Press.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Martin Embley, Associate Editor

Accepted April 14, 2009