# Evaluating Evolutionary Constraint on the Rapidly Evolving Gene matK Using Protein Composition

**Michelle M. Barthet · Khidir W. Hilu**

**Abstract** The rapidly evolving chloroplast *matK* gene has nucleotide and amino acid substitution rates suggestive of progression toward a pseudogene state. However, molecular evidence has demonstrated that *matK* is expressed and functional. We explore in this paper the underlying factors behind the mode and tempo of *matK* evolution that allow this protein coding gene to accommodate such elevated rates of substitution and yet maintain functionality. Conservative amino acid replacement may reconcile the fast evolutionary rate in *matK* with conservation in protein function. Based on this premise, we have examined putative amino acid sequences for MATK from across green plants to determine constraint on this protein as indicated by variation in composition of amino acid side chain category. Amino acids in the MATK ORF were divided into six categories based on chemical properties of their side chains: nonpolar, uncharged (pH 7), basic, acidic, aromatic, and "special" (amino acids that specifically affect protein structure, i.e., proline, glycine, and cysteine). The amount of standard deviation (SD) in side chain composition was used as a measure of variation and constraint, where a low SD implied high evolutionary constraint and a high SD implied low constraint. Further, we used secondary structure prediction to evaluate if conservation observed in side chain composition was reflected in stable predicted structure. The results of this study demonstrate evolutionary constraint on MATK, identify three regions of functional importance, show highly conserved secondary structure, and support the putative function of MATK as a group II intron maturase.

## Introduction

Several hypotheses have been proposed regarding factors that govern evolutionary constraint on protein coding genes (Jukes and Kimura 1984; Graur 1985; Xia and Li 1998; Ophir et al. 1999). One of the most widely accepted hypotheses in protein evolution is that stringency in structural and functional constraint determines the rate of amino acid substitution in a coding sequence (e.g., Jukes and Kimura 1984; Ophir et al. 1999). In other words, proteins (or regions of proteins) with very strict functional and structural requirements will undergo negative selection and have low amino acid substitution rates (Tourasse and Li 2000). The converse is that relatively high amino acid substitution rates in protein-coding genes imply low functional and structural constraints. This loss of constraint could signify a trend toward loss of function and descent into a pseudogene state. The rapidly evolving chloroplast *matK* gene, however, appears to contradict this assumption despite substitution rates three times those of some chloroplast protein-coding genes at the nucleotide level and sixfold at the amino acid level (Olmstead and Palmer 1994; Soltis and Soltis 1998).

The *matK* gene is located in the large single-copy region of the chloroplast genome, nested between the 5' and the 3' exons of *trnK*, tRNA-lysine, within a group II intron. The

M. M. Barthet · K. W. Hilu
Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24061, USA

M. M. Barthet (✉)
School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia
e-mail: michelle.barthet@bio.usyd.edu.au

gene is approximately 1500 bp in length, corresponding to 500 amino acids. Nucleotide variation per site in *matK* is three times higher than in *rbcL* (the large subunit of RU-BISCO) (Soltis and Soltis 1998), and the amino acid substitution rate is six times that of *rbcL* (Olmstead and Palmer 1994). Contributing to the high amino acid substitution rate in *matK* are the almost-equal ratios of substitutions among all three codon positions (Johnson and Soltis 1994; Hilu and Liang 1997; Xiang et al. 1998; Hilu et al. 2003). The high rate of substitution in this gene has resulted in an increased number of parsimony informative sites and strong phylogenetic signal, contributing to its use to discern evolutionary histories at several taxonomic levels (e.g., Johnson and Soltis 1994; Hayashi and Kawano 2000; Hilu et al. 2003; Cameron 2005; Müller et al. 2006). The wealth of phylogenetic information generated from *matK* has made it an extremely valuable gene for systematic and evolutionary studies.

In addition to the importance of *matK* in systematic studies, it is the only putative group II intron maturase encoded in the chloroplast genome (Neuhaus and Link 1987). This putative function is based on the homology of a region in the carboxy terminus of MATK to domain X of mitochondrial group II intron maturases (Neuhaus and Link 1987). Maturases are enzymes that catalyze nonautocatalytic intron removal from premature RNAs. Substrates proposed to require MATK for intron excision include RNA transcripts for the *trnK, trnA*, *trnI*, *rps12*, *rpl2*, and *atpF* genes (Ems et al. 1995; Jenkins et al. 1997; Vogel et al. 1997, 1999). The tRNA or protein products from these genes are required for normal chloroplast function, implicating an essential function for MATK in the chloroplast.

Recent study has demonstrated that *matK* is transcribed and translated into a protein product, which functions in light-regulated activities and plant development (Barthet and Hilu 2007). These findings suggest that functional and structural constraint must exist at some level in this protein to maintain its expression and activity in the plant. We propose that both functional and structural constraint exists on *matK* but these constraints are imposed at higher levels of protein structure rather than amino acid sequence. We hypothesize that the majority of amino acid substitution in MATK is constrained at the structural and biochemical level by chemically conserved amino acid replacement. Chemically conserved amino acid replacement here refers to the substitution of one amino acid for another of similar side chain chemical properties (e.g., acidic for an acidic or basic for a basic amino acid). Because these replacements may not change the polarity or structural framework of the protein, they act as a form of silent mutation (Graur and Li 1988; Kellogg and Juliano 1997; Wolfe and dePamphilis 1998), minimizing the impact of nonsynonymous

substitutions on protein structure and function. Thus, despite the high rate of nonsynonymous substitution in *matK*, most of these substitutions may be chemically conserved and act similarly to silent mutations, not altering protein structure and function.

In this study, we have used the *matK* gene as a model to determine whether conservative amino acid replacement could reconcile fast evolution on a nucleotide level with expressed and functional protein in rapidly evolving protein-coding genes. We examined whether conservative amino acid replacement is occurring in the rapidly evolving protein MATK and then determined if this process is maintaining elements of structure in this protein. To test this hypothesis, we have examined the predicted open reading frame (ORF) of MATK from several green plant species and characterized composition (side chain composition) for the six amino acid categories: nonpolar, uncharged (pH 7), basic, acid, aromatic, and "special" (amino acids that specifically affect protein structure, i.e., proline, glycine, and cysteine). Further, we analyzed variation in side chain composition in order to determine if constraint exists on these six categories. These six amino acid categories represent the main chemical groups by which amino acids are typically divided with reference to structure and function (Zvelebil et al. 1987; Lodish et al. 2000). Further, we identified regions of high functional constraint in MATK, reflecting the core elements required for function of this protein. We also compared side chain composition and amount of deviation in this composition in MATK to those of the slowly evolving protein RBCL, the pseudogene *infA*, and the mitochondrial maturase MATR. This comparison has provided a context for determining the level of functional constraint observed in MATK relative to other proteins in the plant cell that exhibit different modes and tempos of evolution. Further, we have used secondary structure prediction to determine if structure is maintained in MATK regardless of the high amino acid substitution rate. We compared this predicted secondary structure to that of the well-characterized bacterial group II intron maturase LTRA (Matsuura et al. 1997; Saldanha et al. 1999; Blocker et al. 2005; Cui et al. 2004; Rambo and Doudna 2004) in order to determine if the predicted MATK structure is characteristic of proteins with this function.

## Materials and Methods

### Source of Sequence Data

All amino acid sequences were downloaded directly from the protein database of GenBank with the exception of the *matK* ORF from *Huperzia lucida* and *Anthoceros formosae*

and all *infA* protein sequences, as amino acid sequences for these were lacking. Nucleotide sequences for *infA*, as well as the *matK* ORF of *Huperzia lucida* and *Anthoceros formosae*, were downloaded from GenBank and translated into amino acid in MacVector or Accelrys DsGene.

## MATK Side Chain Composition

Amino acid sequence from a total of 68 species from 53 families ranging across green plants was used to assess side chain composition and variability of the MATK ORF (Supplementary Table S1). This data set was split into three smaller data sets (A, B, and C) to simplify examination of side chain composition across the full-length ORF (data set A), among domains and sectors (data set B), and in comparison to data sets with and without indels (data set C). Methods for all analyses and reasons for the inclusion or exclusion of species or plant groups in each data set (A, B, and C) are discussed in the following sections.

## Full-Length MATK ORF Side Chain Composition (Data Set A)

A data set of 58 species from 49 families (data set A; Supplementary Table S1) spanning green plants was used to evaluate MATK amino acid composition using the SAPS program (Brendel et al. 1992) in the San Diego Supercomputer Center's (SDSC) Biology Workbench (version 3.2). Forty-one of these species represent various angiosperm lineages (the largest land plant lineage), while the remaining species represent algae (one), bryophytes (three), monilophytes (three), and gymnosperms (four). Amino acids were classified into the following six categories based on the chemical properties of their side chains: nonpolar (hydrophobic), polar uncharged (pH 7), basic, acid, aromatic, and special. Division of amino acid categories followed that of Lodish et al. (2000) and Zvelebil et al. (1987), with two exceptions; tyrosine was included in the polar uncharged (pH 7) and aromatic categories but not the nonpolar category, and histidine was only included in the basic amino acid category, and not considered aromatic. The classification of tyrosine and histidine conflicted between Lodish et al. (2000) and Zvelebil et al. (1987). Therefore, the categorization of these two amino acids followed a consensus based on Zvelebil et al. (1987), Lodish et al. (2000), and Alberts et al. (2002). The composition of each amino acid category was evaluated as a percentage of total protein, and the standard deviation (SD) of this composition was calculated as a measure of variation.

## MATK Domains and Sectors (Data Set B)

The MATK ORF of 31 green plant species (data set B; Supplementary Table S1) was divided into the two domains, N-terminal region and domain X, for comparison of hypothesized functional versus nonfunctional segments in this protein. Data set B is a subset of data set A. Species used in data set B covered the same evolutionary range as data set A and included species from green algae (*Chaetosphaeridium*) to angiosperms. However, the number of angiosperm species examined was reduced to lessen the weight given to this particular plant lineage, and the bryophytes *Marchantia* and *Anthoceros* were excluded. *Anthoceros* was excluded because the MATK ORF from this bryophyte contained several premature stop codons, which may bias analysis of different regions of MATK. Several amino acids were missing from the N-terminal region of the MATK ORF from *Marchantia*, thus this bryophyte was also eliminated from data set B. Boundaries of each domain were determined by Pfam (Bateman et al. 2002). The Pfam program uses sequences stored in GenBank to generate multiple sequence alignments of an entered sequence of interest and identifies regions homologous to known protein family domains. The MATK ORF was further subdivided into seven sectors since division into only the two large domains of the N-terminal region and domain X may overlook smaller regions of functional constraint and result in overestimation of variation in side chain composition. Each sector had 72 amino acids on average, with the exception of the carboxy terminus sector, which had an average of 80 amino acids. Because the MATK ORF contains several indels that can adjust the position of otherwise similar sequence elements, boundaries of sectors were determined by regions of conserved sequence on both sides of the sector. These conserved sequence elements were identified by alignment of the MATK ORF from species in data set B using MacVector and Accelrys DsGene. Validity of this amino acid alignment was determined by phylogenetic parsimony analysis using PAUP* (version 4.0b6; Swofford 2001) with a 50% majority-rule heuristic search strategy with stepwise addition. To provide a stringent test of the alignment, taxon sampling included all species in data set A in tree reconstruction. Bootstrap values for trees were calculated using 500 replicates. Gaps were treated as missing data and all aligned positions were given equal weight. *Chaetospiridium* was used as the outgroup taxon. A consensus tree was compiled and compared to the land plant phylogeny produced by Magallón and Sanderson (2002). The program MEME (Bailey and Elkan 1994) was used to further confirm sector boundaries identified by alignment. MEME is a program designed specifically to find conserved motifs in unaligned sequences and, therefore, presents a completely

unbiased identification of conserved amino acid sequences in the MATK ORF.

Side chain composition and variability for the domains and sectors were determined as previously stated for the entire MATK ORF. Side chain compositions of each domain or sector within MATK were compared against each other using a Student $t$-test in Excel. Since the side chain composition for each chemical group is a percentage of the total protein, data were transformed using arcsin square root transformation prior to statistical comparison. Because amino acids defined for the aromatic and "special" amino acid categories overlap those of the nonpolar and uncharged (pH 7) side chain categories, only the four amino acid categories of nonpolar, uncharged (pH 7), basic, and acidic amino acids were included for SD comparisons between domains and sectors of MATK.

## Examination of the Impact of Indels on Composition (Data Set C)

An alignment of sequences from 14 seed plant species (angiosperms and gymnosperms: data set C; Supplementary Table S1) that contained only a maximum of 2-amino acid indels was analyzed in the same manner as described previously and compared to the results from data sets A and B to determine the impact of indels on estimation of side chain composition and variability in the MATK ORF. Only seed plants were used in this part of the study because inclusion of taxa beyond this group greatly increased the number and size of indels present in this protein.

## Comparison of the MATK ORF to That of Other Proteins

The amount of functional constraint in MATK (high, low, or intermediate) was determined by comparison of variation in side chain composition in the MATK ORF to the pseudogene *infA* (Millen et al. 2001), the slow-evolving functional protein RBCL (Wolfe et al. 1987; Chase et al. 1993; Kellogg and Juliano 1997), and the mitochondrial maturase MATR (Farré and Araya 1999). Pseudogenes, such as *infA*, lack selective constraint (Echols et al. 2002), while the conserved protein RBCL has very high functional constraint (Albert et al. 1994; Kellogg and Juliano 1997), presenting two extremes of constraint for comparison. Comparison to MATR, another group II intron maturase, provided a standard to determine if the level of variation observed in MATK is normal for this particular enzyme family. Seven angiosperm species from two families were used to compare variation in side chain composition between INFA and MATK (Supplementary Table S2). The chloroplast *infA* gene has been horizontally transferred to the nucleus in several plants, but a residual pseudogene copy has been retained in several chloroplast genomes (Millen et al. 2001). Only the pseudogene copy of *infA* was used in this analysis. Protein sequences for this pseudogene were generated by in silico translation in Accelrys DsGene and included premature stop codons. Thirty-two plant species from 31 families across land plants were used to compare MATK and MATR side chain composition and variability (Supplementary Table S3). Comparisons between MATK and RBCL were based on sequences for 21 species from 16 green plant families (Supplementary Table S4). Selection of species for each comparison was determined by availability in GenBank of complete sequences for each protein in the exact same species. Side chain composition and variability were determined as described previously for the MATK ORF.

## MATK Secondary Structure Prediction

The secondary structure of MATK was predicted using the PredictProtein server (Rost and Sander 1993; Rost et al. 2004) and JPRED (Cuff et al. 1998; Cuff and Barton 2000). Methodology of prediction from both servers includes input of sequence data into PSI-BLAST to scan the SWISS/PROT/TRMBL database for homologous sequences and generate a multiple sequence alignment. The PredictProtein server also utilizes MaxHom (Sander and Schneider 1991) to generate a second alignment profile. In the JPRED program, alignment profiles are input into Jnet neural networks trained on several hundred known protein structures, to generate the predicted secondary structure. JPRED prediction is reported as 76.4% accurate (Cuff and Barton 2000). The PredictProtein server contains the two secondary structure prediction programs PHD and PROF. PROF predictions have been shown to be more accurate than those from PHD (Rost 2001; Rost et al. 2004). Only PROF predictions, therefore, were used in the current study. PROF predictions employed in this study were those generated with the addition of a position-specific reliability index, shown to increase accuracy in the prediction to 82% (Rost and Sander 1993). Input sequences used for secondary structure prediction include the MATK ORF from *Oryza sativa*, *Arabidopsis thaliana, Pinus korainsis* (Supplementary Table S1), and the group II intron maturase LTRA (GenBank accession no. P0A3U0). The evolutionary breadth of species used for comparison of predicted MATK secondary structures allowed comparison of conserved structural elements between monocots and eudicots as well as among angiosperms and gymnosperms. Although predictions from PROF are reported to be more accurate than those of JPRED (Rost 2001), secondary structure prediction was performed for the MATK ORF from *Oryza sativa* using both PROF and JPED in order to compare the accuracy of PROF to another prediction program exterior to this server.

Potential for transmembrane helices was determined using TMAP and TMHMM transmembrane prediction computer programs (Persson and Argos 1994) in SDSC's Biology Workbench. Transmembrane regions were analyzed for the MATK ORF from all species in data set A as well as for the group II intron maturases MATR (GenBank accession no. AE47664), COB-I1 (GenBank accession no. X54421), COX1-I2 (GenBank accession nos. NC_005256 and CAC28096), and LTRA.

## Results

### MATK Side Chain Composition

#### Full-length ORF (data set A)

Examination of the putative MATK reading frame from species in data set A revealed an average of 47% ± 1.6% nonpolar (hydrophobic), 28% ± 1.7% uncharged (pH 7),

17% ± 1.2% basic, and 8% ± 0.9% acidic amino acids (Table 1, Fig. 1A). When these amino acid groups were divided further, it was observed that MATK had an average of 15% ± 2.0% aromatic and 8% ± 1.0% "special" (proline, glycine, and cysteine) amino acids. This composition was fairly uniform across all green plant taxa examined, with the exception of the Gnetophyta (Fig. 1B), which displayed an average 12.5% increase in nonpolar amino acids compared to other plant taxa. For all species examined, nonpolar amino acids constituted the largest percentage of any type of amino acid in MATK (47% ± 1.6%), while acidic and "special" amino acids comprised the smallest proportion (8% ± 0.9% and 8% ± 1.0%, respectively) (Table 1).

The considerable increase in nonpolar amino acids in the two Gnetophyta species *Gnetum gnemon* and *Welwitschia mirabilis* was paralleled by a 25% increase in aromatic amino acid content (Fig. 1B). To test if the hydrophobic content observed for the Gnetophyta was significantly higher than that for other plant genera, we increased the number of Gnetophyta species examined to six by adding *Ephedra*

**Table 1** Side chain composition of MATK

| Region | Amino acid category (% ± SD) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Nonpolar | Uncharged | Basic | Acid | Aromatic | "Special" | Average SD |
| MATK ORF | 47 ± 1.65 | 28 ± 1.64 | 17 ± 1.25 | 8 ± 0.84 | 15 ± 1.93 | 8 ± 1.05 | ±1.35 |
| | *47 ± 0.79* | *28 ± 1.35* | *17 ± 0.98* | *8 ± 0.74* | *15 ± 1.76* | *8 ± 0.95* | *±0.96* |
| N-terminal region | 47 ± 2.32 | 28 ± 2.23 | 16 ± 1.61 | 8 ± 0.91 | 16 ± 2.25 | 8 ± 1.44 | ±1.77 |
| | *47 ± 1.26* | *29 ± 1.50* | *16 ± 1.15* | *8 ± 0.80* | *16 ± 1.94* | *7 ± 1.12* | *±1.18* |
| Domain X | 46 ± 1.84 | 21 ± 3.31 | 21 ± 1.82 | 8 ± 1.55 | 13 ± 0.59 | 10 ± 1.56 | ±2.13 |
| | *46 ± 1.51* | *24 ± 1.98* | *22 ± 1.63* | *8 ± 0.78* | *11 ± 1.76* | *10 ± 0.98* | *±1.48* |
| Sector 1 | 42 ± 4.17 | 31 ± 3.43 | 17 ± 3.72 | 10 ± 2.80 | 15 ± 3.94 | 7 ± 2.91 | ±3.53 |
| | *42 ± 2.54* | *31 ± 2.40* | *16 ± 3.52* | *11 ± 3.31* | *14 ± 2.48* | *7 ± 2.10* | *±2.94* |
| Sector 2 | 47 ± 2.91 | 28 ± 3.87 | 13 ± 2.26 | 11 ± 2.22 | 13 ± 3.01 | 8 ± 1.96 | ±2.81 |
| | *48 ± 1.80* | *30 ± 1.64* | *13 ± 3.34* | *10 ± 2.62* | *12 ± 3.12* | *8 ± 1.88* | *±2.35* |
| Sector 3 | 52 ± 4.51 | 24 ± 3.54 | 18 ± 2.62 | 6 ± 1.23 | 16 ± 4.26 | 7 ± 2.97 | ±2.97 |
| | *51 ± 1.69* | *25 ± 0.70* | *18 ± 2.19* | *6 ± 2.19* | *17 ± 4.57* | *6 ± 2.02* | *±1.69* |
| Sector 4 | 44 ± 3.61 | 29 ± 3.21 | 19 ± 3.98 | 8 ± 1.74 | 20 ± 3.03 | 7 ± 2.69 | ±3.13 |
| | *44 ± 3.81* | *29 ± 1.65* | *18 ± 1.90* | *9 ± 3.39* | *20 ± 2.70* | *8 ± 2.84* | *±2.69* |
| Sector 5 | 51 ± 3.45 | 30 ± 4.18 | 15 ± 2.55 | 3 ± 1.29 | 19 ± 4.38 | 9 ± 2.73 | ±2.87 |
| | *50 ± 2.37* | *31 ± 1.10* | *15 ± 3.59* | *3 ± 2.02* | *17 ± 3.30* | *8 ± 1.95* | *±2.27* |
| Sector 6 | 48 ± 2.82 | 25 ± 3.10 | 18 ± 2.86 | 9 ± 2.72 | 11 ± 3.27 | 13 ± 1.71 | ±2.88 |
| | *48 ± 2.40* | *25 ± 1.99* | *19 ± 2.86* | *8 ± 3.11* | *10 ± 2.35* | *13 ± 1.31* | *±2.59* |
| Sector 7 | 46 ± 3.11 | 24 ± 4.86 | 20 ± 2.92 | 9 ± 2.43 | 12 ± 3.52 | 7 ± 2.08 | ±3.33 |
| | *47 ± 2.41* | *23 ± 1.28* | *20 ± 4.80* | *10 ± 2.98* | *11 ± 2.55* | *7 ± 2.17* | *±2.87* |

*Note.* "Uncharged" refers to amino acids uncharged at pH 7. Categories consisted of the following amino acids: nonpolar (A, G, V, L, I, P, F, M, W, C); uncharged, pH 7 (N, Q, S, T, Y); basic (K, R, H); acidic (D, E); aromatic (F, W, Y); and "special" (C, G, P). Side chain composition and corresponding SD for data set C (species with few indels) are noted in italics. Side chain composition and SD for the full-length MATK ORF from species in data set A is noted as "MATK ORF." Composition and SD of each domain and sector were determined using species from data set B. Average SD is the average standard deviation in the four amino acid categories of nonpolar, uncharged (pH 7), basic, and acidic. Aromatic and "special" amino acids were not included in the average SD calculation to avoid overestimation of SD due to redundancy of amino acids among these two groups and that of the other four amino acid categories

**Fig. 1** Average side chain composition of the full-length MATK ORF (data set A). Amino acids were categorized as nonpolar, uncharged polar at pH 7, basic, acid, aromatic, or "special." Categories consisted of the following amino acids: nonpolar (A, G, V, L, I, P, F, M, W, C); uncharged, pH 7 (N, Q, S, T, Y); basic (K, R, H); acidic (D, E); aromatic (F, W, Y); and "special" (C, G, P). Standard error is shown. **A** Average percentage of each amino acid category in total protein. **B** Side chain composition of MATK across representative green plant genera. Taxa are arranged in phylogenetic order from left (basal lineages) to right (recently diverged lineages)
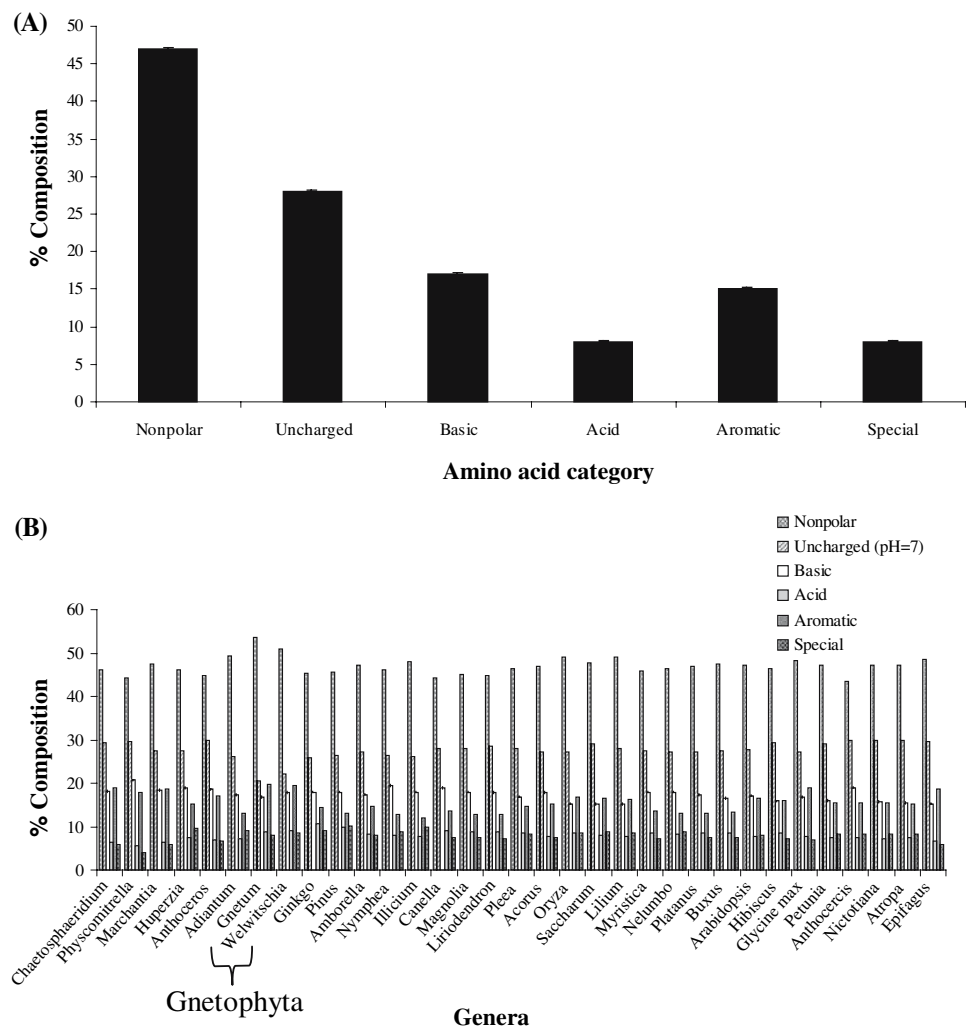


*sinica, Ephedra sarcocarpa, Gnetum africanum,* and *Gnetum gnemonoides* (Supplementary Table S1). We then compared the nonpolar amino acid composition of these six gnetophytes to an equal number of (1) other gymnosperm species (*Pinus korainsis, Cryptomeria japonica, Phyllocladus trichomanoides, Taxus cuspidate, Sequiadendron giganteum, Ginkgo biloba*; Supplementary Table S1), (2) bryophytes (excluding *Marchantia*) and monilophytes (data set A), and (3) basal angiosperms (the first six angiosperms from data set A; Supplementary Table S1). This comparison revealed that the nonpolar content of MATK protein in gnetophytes was significantly higher than that from all three groups examined (Student's $t$-test, $t_{10} = 10.9$, 6.4, and 7.3, respectively; $p < 0.0001$ for all comparisons). These results suggest a molecular evolutionary shift in MATK composition in this one plant lineage.

### Domains and sectors (data set B)

Division of the MATK ORF into the two major domains, N-terminal region and domain X, or separation into

seven sectors revealed a similar pattern of composition as seen with the full-length ORF (Table 1, Fig. 2). Although the overall proportions of side chain categories were similar across domains and sectors, there were several differences in chemical amino acid makeup that may be relevant to the specific structure and function of each region. Domain X had a significantly higher amount of basic amino acids ($21\% \pm 1.8\%$) compared to the N-terminal region ($16\% \pm 1.6\%$; Student's $t$-test, $t_{58} = 10.5$, $p < 0.0001$). The percentage of "special" amino acids, which directly affect elements of protein structure, were also statistically higher in domain X ($10\% \pm 1.5\%$) than in the N-terminal region ($8\% \pm 1.4\%$; $t_{58} = -6.6$, $p < 0.0001$), suggesting the existence of a complex structure for this domain.

Evaluation of the seven sectors revealed two large peaks of nonpolar amino acids in sectors 3 and 5 ($52\% \pm 4.51\%$ and $51\% \pm 3.45\%$, respectively) (Table 1, Fig. 2C). These peaks could form hydrophobic pockets in MATK protein structure. The distribution of aromatic and "special" amino acids varied among sectors. Sector 6, which overlaps
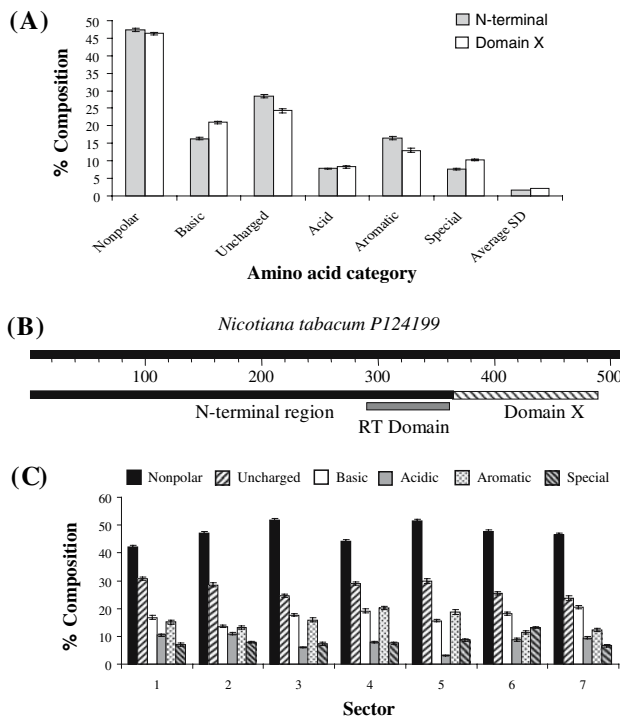
**Fig. 2** Comparison of side chain composition between the two MATK domains, N-terminal and domain X (data set B). **A** Comparison of side chain composition between the N-terminal region and domain X. **B** Schematic of the relative positions of the N-terminal region, RT domain, and domain X to the full-length MATK ORF. **C** Side chain composition of each sector. Standard error is shown. Amino acid categories are defined in Fig. 1

domain X, had the highest proportion of "special" amino acids (13% ± 1.7%) (Fig. 2C).

## Constraint on MATK

We evaluated functional constraint on MATK protein evolution by calculating the SD in side chain composition, where a high SD would indicate less constraint, and a low SD would indicate high constraint. Amount of acidic amino acids (SD, ± 0.89) in MATK was highly invariable across all green plants studied (Table 1), demonstrating high evolutionary constraint on this amino acid group compared to the other amino acid categories. In contrast, the composition of nonpolar (SD, ± 1.65) and uncharged (pH 7) amino acids (SD, ± 1.64) was the least constrained.

Although domain X is the putative functional domain for MATK maturase activity (Neuhaus and Link 1987; Mohr et al. 1993) and would be expected to have higher functional constraint than the N-terminal region, no significant difference in average SD of side chain composition was observed between the two domains (Student's $t$-test,

$t_{10} = 0.92$, $p = 0.18$). In fact, comparison of the average SD for the amino acid categories nonpolar, uncharged (pH 7), basic, and acid between the two domains showed that the N-terminal region had less variation in side chain composition than domain X (SD, ± 1.77 and ± 2.13, respectively; Table 1).

Analysis of variation in side chain composition for each of the seven sectors revealed three regions of high functional constraint (low standard deviation) and two "hot spots" of variation (SD, ± 3.53 and ± 3.33 for sectors 1 and 7, respectively). Sectors 2, 5, and 6 had the lowest variation in side chain composition (Fig. 3A) and had similar SDs (± 2.8, ± 3.0, and ± 2.9, respectively). Sectors 5 and 6 correspond to functional domains previously identified by sequence comparison (Mohr et al. 1993). Sector 5 overlaps the remaining elements of the reverse-transcriptase (RT) domain found in other group II intron maturases (Fig. 3B and C), while sector 6 overlaps domain X.

Sector 2 does not overlap any previously determined functional domains for MATK. Therefore, Pfam analysis was conducted to determine similarity to known protein domains. Pfam analysis was performed using the Pfam gathering threshold or E-values from 1.0 to 10.0 on 10 taxa from data set A (noted by an asterisk in Supplementary Table S1), which represent land plants from bryophytes to angiosperms. No significant similarity to known protein domains was identified for this sector by Pfam analysis other than as part of the N-terminal region of MATK.

## Impact of MATK Indels (Data Set C)

Compared across green plants, the MATK ORF contains several indels that may bias our analyses of side chain composition. We analyzed, therefore, an alignment of 14 taxa (data set C), containing a minimal number of indels, for percentage side chain composition and SD of each amino acid category for the entire MATK ORF, the two domains, and the seven sectors and compared with corresponding results from data sets A (overall ORF) and B (domains and sectors). The percentages of each side chain category for the entire ORF, the two domains, and the seven sectors were not statistically different among data sets C and A or B (Table 1). SD was lower in almost all cases for data set C compared to the larger data sets A and B (Table 1), but these differences were not statistically significant and are most likely due to differences in evolutionary breadth of taxa sampled. The 14 taxa in data set C included 13 angiosperms and 1 gymnosperm, compared to a very broad sampling of green plants for data sets A and B. To examine the impact of data sampling, species in data sets A and B were separated into more closely related phylogenetic groups (bryophytes and monilophytes, gymnosperms, and angiosperms) and reassessed for SD in
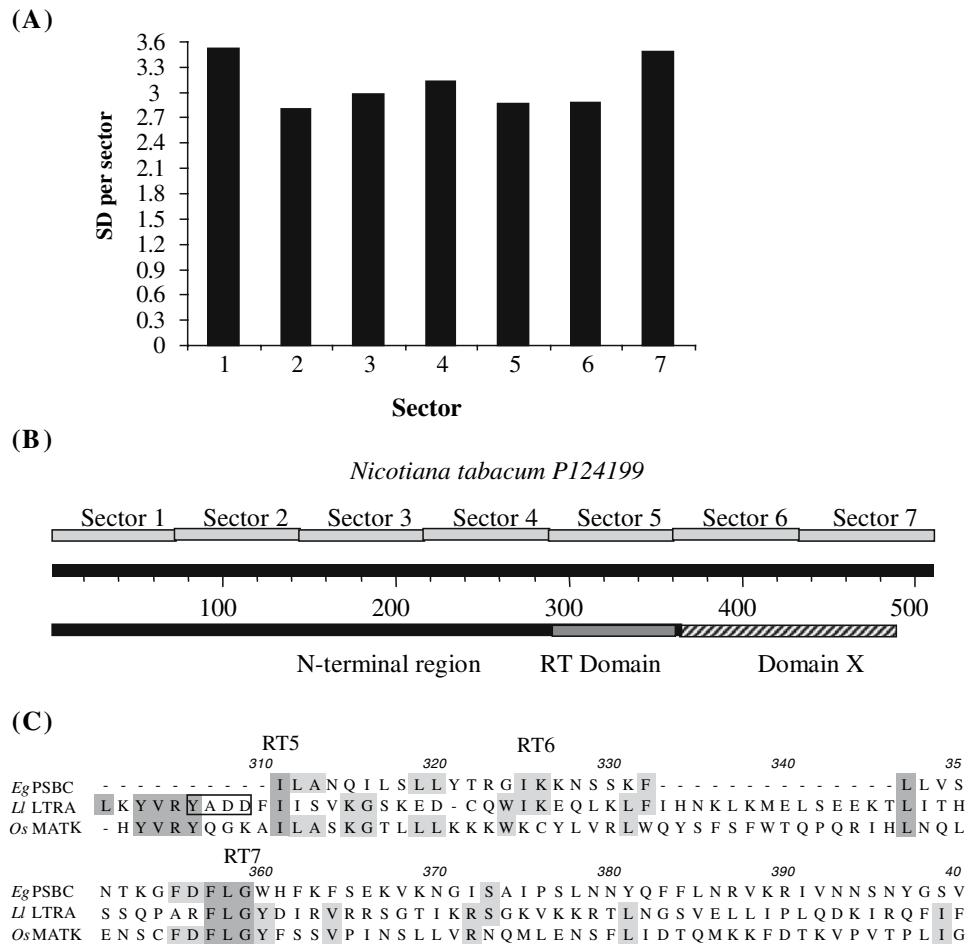
**(A)**



**(B)**

*Nicotiana tabacum P124199*



**(C)**



**Fig. 3** Evaluation of variation in side chain composition for the seven sectors of the MATK ORF. **A** Average SD of the four amino acid categories of nonpolar, acidic, basic, and uncharged (pH 7) for each sector. **B** Schematic representation of the division of the MATK ORF into the seven sectors, with the relative position of the N-terminal region, RT domain, and domain X plotted. The positions of the N-terminal region and domain X were determined by Blastp search in GenBank. The location of the RT domain was based on consensus sequence determined by Mohr et al. (1993). **C** Alignment of the homologous sequence blocks of the RT domain in *Euglena gracilis* (Eg) PSBC (ORF located in *psbC* intron 4), *Lactococcus lactis* (Ll) LTRA group II intron maturase, and *O. sativa* (Os; rice) MATK. Sequence blocks are noted as 5, 6, and 7 following Mohr et al. (1993). The YXDD conserved sequence found in the majority of functional reverse transcriptases is boxed. One hundred percent identity in sequence is highlighted in dark gray, consensus match is highlighted in light gray, and mismatched amino acids are in white

side chain composition. The SD for each amino acid category was lower when comparing the angiosperm subset with the full data set, but higher in some categories when comparing the bryophyte and gymnosperm subsets (data not shown). Therefore, we conclude that indels did not impact our assessment of side chain composition or SD in MATK and that the small difference in SD of side chain composition observed with data set C can be attributed to taxon sampling.

Side Chain Composition in MATK Versus INFA, RBCL, and MATR

In order to determine if the amount of SD observed in MATK side chain composition was an indication of functional constraint, this SD was compared to those of the translated pseudogene *infA*, the highly conserved protein RBCL, and the mitochondrial maturase MATR. Our analysis showed that these three proteins did not differ in relative side chain composition from MATK (Table 2). The four proteins did, however, differ considerably in SD of side chain composition. SD in side chain composition from translated *infA* pseudogene sequences was significantly higher than that of MATK (SD, $\pm$ 2.81 and $\pm$ 0.75, respectively; Student's $t$-test, $t_6 = -2.9$, $p = 0.027$) (Table 3). No significant difference was found in variation of side chain composition between MATK and the mitochondrial maturase MATR (SD, $\pm$ 1.75 and $\pm$ 0.98, respectively; $t_6 = -1.5$, $p = 0.195$) (Table 2). MATK, however, had a significantly higher amount of SD in side chain

**Table 2** Comparison of average percentage side chain composition ± SD of MATK to that of INFA, RBCL, and MATR

| Amino acid category | Protein | | | | | |
|---|---|---|---|---|---|---|
| | MATK[1] | INFA[1] | MATK[2] | MATR[2] | MATK[3] | RBCL[3] |
| Nonpolar (%) | 47 ± 0.25 | 51 ± 4.81 | 41 ± 2.90 | 51 ± 1.56 | 47 ± 1.34 | 52 ± 0.65 |
| Uncharged (%) | 29 ± 1.33 | 25 ± 1.91 | 27 ± 1.78 | 22 ± 1.21 | 28 ± 1.48 | 19 ± 0.58 |
| Basic (%) | 16 ± 0.65 | 16 ± 2.17 | 23 ± 1.69 | 19 ± 0.73 | 17 ± 1.55 | 16 ± 0.42 |
| Acid (%) | 8 ± 0.77 | 8 ± 2.37 | 8 ± 0.63 | 9 ± 0.42 | 8 ± 1.14 | 13 ± 0.21 |
| Aromatic (%) | 15 ± 0.31 | 11 ± 2.04 | 14 ± 1.83 | 8 ± 0.63 | 15 ± 2.13 | 9 ± 0.16 |
| "Special" (%) | 8 ± 0.27 | 11 ± 3.08 | 8 ± 0.77 | 14 ± 1.35 | 8 ± 1.47 | 16 ± 0.39 |
| Average SD | 0.75 | 2.81 | 1.75 | 0.98 | 1.38 | 0.46 |

*Note.* Superscripts refer to different data sets used for comparisons of MATK to [1]INFA (Supplementary Table S2), [2]MATR (Supplementary Table S3), and [3]RBCL (Supplementary Table S4). "Uncharged" refers to amino acids uncharged at pH 7. Amino acid categories are defined in the Note to Table 1

composition than RBCL (SD, ± 1.38 and ± 0.46, respectively; $t_6 = -.8$, $p = 0.001$).

## MATK Structural Elements

A crystal structure does not exist for any group II intron maturase. We, therefore, used secondary structure prediction to determine if MATK has conserved structural elements across land plants and if those elements show homology to other group II intron maturases, specifically the *Lactococcus lactis* group II intron maturase LTRA. We examined predicted secondary structure for three model species (*Arabidopsis thaliana*, *Oryza sativa*, and *Pinus koriansis*). We compared predicted secondary structure for *O. sativa* from two different programs (JPRED and Prof from the Predict Protein server) to determine the reliability of these structural predictions. Overall predictions from both programs were in agreement, although slight discrepancies in number and size of a few helices were noticeable (data not shown). Alignments shown in Figs. 4 and 5 were generated from PROF predictions, which are considered slightly more accurate than JPRED predictions (Rost 2001). The positions of nearly all structural elements in MATK were identical for all three species examined except for two predicted α-helical regions and a few loop and turn regions (Fig. 4), supporting a conserved structure for MATK.

Alignment of the MATK predicted secondary structure from rice to that of the bacterial group II intron maturase LTRA identified several regions of highly conserved structure despite divergence in amino acid sequences (Fig. 5). Structural homology was found in the N-terminal region of MATK to sequence blocks RT0, RT3, RT5, RT6, and RT7 and insertion 3a of the LTRA RT domain previously aligned to the RT domain of HIV-1 by Blocker et al. (2005). Further, three α-helices in domain X of MATK correspond to helices αH, αI, and αJ in domain X of LTRA (Fig. 5). MATK was

also found to contain a region homologous to the "ti" insertion found between helix αH and helix αI of domain X of LTRA (Blocker et al. 2005) (Fig. 5).

Since hydrophobic segments in protein structure could indicate transmembrane regions, we analyzed the MATK reading frame from all taxa in data set A using the TMHMM
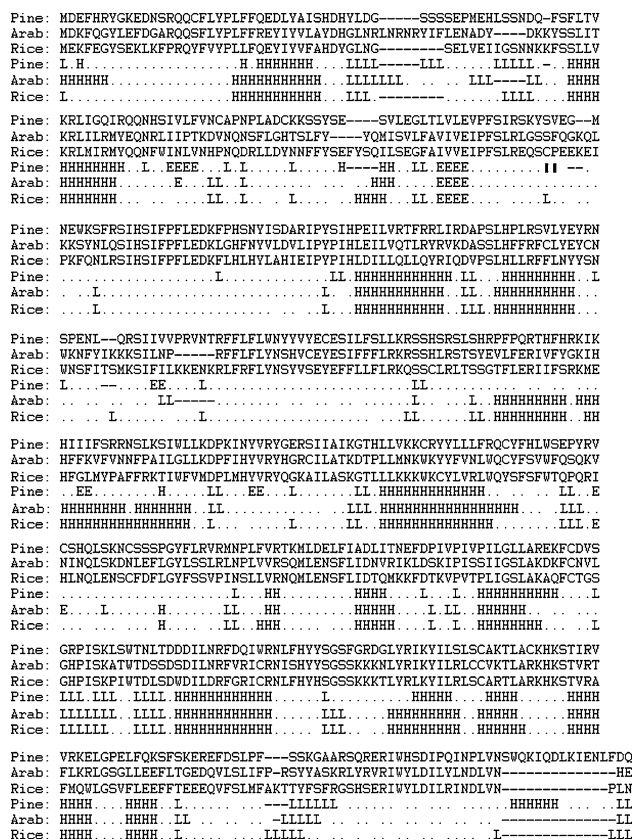
```
Pine: MDEFHRYGKEDNSRQQCFLYPLFFQEDLYAISHDHYLDG-----SSSSEPMEHLSSNDQ-FSFLTV
Arab: MDKFQGYLEFDGARQQSFLYPLFFREYIYVLAYDHGLNRLNRNRYIFLENADY----DKKYSSLIT
Rice: MEKFEGYSEKLKFPRQYFVYPLLFQEYIYVFAHDYGLNG--------SELVEIIGSNNKKFSSLLV
Pine: L.H..................H.HHHHHHH....LLLL-----LLL.......LLLLL.-..HHHH
Arab: HHHHHH...............HHHHHHHHHH...LLLLLLL. . . . LLL-----LL..HHHH
Rice: L..................HHHHHHHHHHH...LLL.--------.....LLLL....HHHH

Pine: KRLIGQIRQQNHSIVLFVNCAPNPLADCKKSSYSE----SVLEGLTLVLEVPFSIRSKYSVEG--M
Arab: KRLILRMYEQNRLIIPTKDVNQNSFLGHTSLFY----YQMISVLFAVIVEIPFSLRLGGSFQGKQL
Rice: KRLMIRMYQQNFWINLVNHPNQDRLLDYNNFFYSEFYSQIILSEGFAIVVEIPFSLREQSCPEEKEI
Pine: HHHHHHH..L..EEEE...L.L.....L.....H----HH..LL.EEEE........II --.
Arab: HHHHHHH.......E...IL..L.......... ..HHH....EEEE..........
Rice: HHHHHHHH... ...IL..L.. . I..... .HHHH...LL.EEEE.... ...L.. ..

Pine: NEWKSFRSIHSIFPFLEDKFPHSNYISDARIPYSIHPEILVRTFRRLIRDAPSLHPLRSVLYEYRN
Arab: KKSYNLQSIHSIFPFLEDKLGHFNYVLDVLIPYPIHLEILVQTLRYRVKDASSLHFFRFCLYEYCN
Rice: PKFQNLRSIHSIFPFLEDKFLHLHVLAHIEIPYPIHLDILLQLLQYRIQDVPSLHLLRFFLNYYSN
Pine: ...................L...............LL.HHHHHHHHHHH.LL...HHHHHHHHH..L
Arab: ...L....................................L..HHHHHHHHHH..LL..HHHHHHHHHH...
Rice: ....L..................... ..L..HHHHHHHHHHH..LLL.HHHHHHHHHH...

Pine: SPENL--QRSIIVVPRVNTRFFLFLWNYYVYECESILFSLLKRSSHSRSLSHRPFPQRTHFHRKIK
Arab: WKNFYIKKKSIILNP-----RFFLFLYNSHVCEYESIFFFLRKRSSHLRSTSYEVLFERIVFYGKIH
Rice: WNSFITSMKSIFILKKENKRLFRFLYNSYVSEYEFFLLFLRKQSSCLRLTSSGTFLERIIFSRKME
Pine: L...--...EE....L..................LL..............
Arab: .. .. . LL-----...............................L. ...L..HHHHHHHHH. HHH
Rice: .. .. L...... ...L........................LL......LL.HHHHHHHHH..HH

Pine: HIIIFSRRNSLKSIWLLKDPKINYVRYGERSIIAIKGTHLLVKKCRYYLLLFRQCYFHLWSEPYRV
Arab: HFFKVFVNNFPAILGLLKDPFIHYVRYHGRCILATKDTPLLMNKWKYYFVNLWQCYFSVWFQSQKV
Rice: HFGLMYPAFFRKTIWFVMDPLMHVVRYQGKAILASKGTLLLKKKWKCYLVRLWQYSFSFWTQPQRI
Pine: .EE........H.....IL...EE....I.......LL.HHHHHHHHHHHH..... .. ...HH.E
Arab: HHHHHHHH.HHHHHHH..LL...........LLL.HHHHHHHHHHHHHHHHH....LLL..
Rice: HHHHHHHH........L.........L..........LL.HHHHHHHHHHHHH...........LLL.E

Pine: CSHQLSKNCSSSPGYFLRVRMNPLFVRTKMLDELFIADLITNEFDPIVPIVPILGLLAREKFCDVS
Arab: NINQLSKDNLEFLGYLSSLRLNPLVVRSQMLENSFLIDNVRIKLDSKIPISSIIGSLAKDKFCNVL
Rice: HLNQLENSCFDFLGYFSSVPINSLLVRNQMLENSFLIDTQMKKFDTKVPVTPLIGSLAKAQFCTGS
Pine: ................. .......L..HH........HHHH...L..L.HHHHHHHHHH....L
Arab: E...L........H....IL...HH..........HHHHH...L.LL.HHHHHH.. ...L
Rice: ..... ......H......IL..HHH.......HHHH...L..L.HHHHHHHHH....L

Pine: GRPISKLSWTNLTDDDILNRFDQIWRNLFHYYSGSFGRDGLYRIKYIILSLSCAKTLACKHKSTIRV
Arab: GHPISKATWTDSSDSDILNRFVRICRNISHYYSGSSKKKNLYRIKYIILRLCCVKTLARKHKSTVRT
Rice: GHPISKPIWTDLSDWDILDRFGRICRNLFHYHSGSSKKKTLYRLKYIILRISCARTLARKHKSTVRA
Pine: LLL.LLL..LLLL.HHHHHHHHHHH......L..........HHHH....HHHH.....HHHH
Arab: LLLLLLL..LLLL.HHHHHHHHHHHH.......LLL...HHHHHHHHH..HHHH......HHHH
Rice: LLLLLL..LLLL.HHHHHHHHHHH.....LLL..HHHHHHHHHH..HHHHH....HHHH

Pine: VRKELGPELFQKSFSKEREFDSLPF---SSKGAARSQRERIWHSDIPQINPLVNSWQKIQDLKIENLFDQ
Arab: FLKRLGSGLLEEFLTGEDQVLSLIFP-RSYYASKRLYRVRIWYLDILYLNDLVN------------HE
Rice: FMQWLGSVFLEEFFTEEEQVFSLMFAKTTYFSFRGSHSERIWYLDILRINDLVN------------PLN
Pine: HHHH....HHHH.L....---LLLLLL .....  . .....HHHHH .... ..LL
Arab: HHHH....HHHH..LL. ... .  -LLLLL .  . . .   --------------LL
Rice: HHHH....HHHH..L.......LLLLL. . . . . .  .L-------------LLL
```

**Fig. 4** Secondary structure alignment of the MATK ORF from pine, *Arabidopsis* (Arab), and rice. Secondary structures shown are those predicted by PROF on the Predict Protein server. H, helices; L, loops; E, B-pleated sheets. Gaps in alignment are noted with a dash
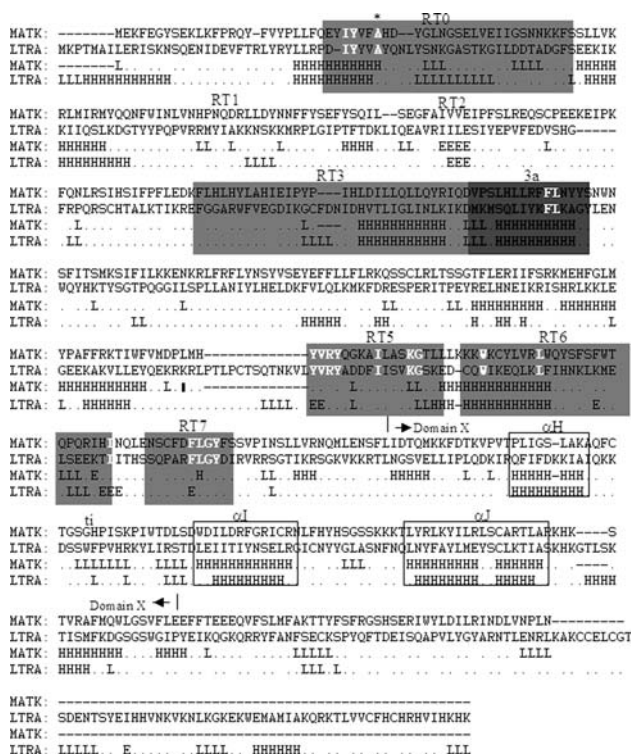
**Fig. 5** Comparison of secondary structure of the rice MATK ORF to LTRA. Gray-shaded boxes denote regions of sequence and structure homology to the RT domain from LTRA. Conserved sequence is highlighted in bold white. Sequence blocks of the RT domain are noted as RT0, RT1, RT2, RT3, RT4, RT5, RT6, and RT7 as delineated by Blocker et al. (2005). The dark-gray-shaded box denotes insertion 3a. Open boxes surround conserved α-helices αH, αI, and αJ of domain X. The "ti" insertion is noted. A line with an arrow delineates where domain X begins and ends relative to the ORFs

and TMAP transmembrane prediction programs (Persson and Argos 1994). We identified one to six putative transmembrane regions in all MATK protein sequences analyzed with the exception of the whisk fern, *Psilotum nudum*, which lacked putative transmembrane segments (data not shown). For comparison, we used these same programs to predict transmembrane regions in protein sequences from other group II intron maturases. Putative transmembrane domains were found in the ORF of the group II intron maturase MATR from *Triticum aesitivum* (one domain), Cob-I1 from *Schizosaccharomyces pombe* (nine domains), Cox1-I2 from *Yarrowia lipolytica* (two domains), and *Penicillium marneffei* (one domain). No transmembrane segments, however, were identified in the ORF of LTRA.

## Discussion

The inherent high rate of nucleotide and amino acid substitution in *matK* is expected to negatively impact its

protein structure and function. Experimental evidence, however, has demonstrated that MATK is expressed and functional in the plant (Vogel et al. 1999; Barthet and Hilu 2007). These findings suggest that constraint exists at some level on MATK to maintain its functional properties. Our analysis of conservative amino acid replacement has demonstrated that, despite its unusual mode and tempo of evolution, MATK is under functional and structural constraint. Further, we have shown that by using SD in side chain composition, we can identify putative functional domains for molecular analysis not easily recognized by sequence similarity alone.

## Composition and Structure

The strong similarity in biochemical composition and structural elements between MATK and LTRA supports the predicted function of MATK as a group II intron maturase. The LTRA group II intron maturase has been well studied for its maturase activity (Matsuura et al. 1997; Saldanha et al. 1999; Blocker et al. 2005; Cui et al. 2004; Rambo and Doudna 2004). Some of the structural and biochemical elements required for LTRA maturase activity include several highly conserved α-helices that form a homodimer when assembled on intron RNA (Saldanha et al. 1999; Rambo and Doudna 2004; Blocker et al. 2005). The MATK ORF is highly hydrophobic (Fig. 1A). Secondary structure prediction demonstrated that this high hydrophobicity is packed into several α-helical regions similar in position to those of LTRA (Fig. 5), suggesting similarity in structure between these two proteins. Furthermore, the template-primer binding tract in LTRA is described as identifiable by a "ribbon of positive charge" and is composed of regions RT2, RT4, RT5, and RT7 of the RT domain and the helices αH and αI of domain X (Blocker et al. 2005). This positive charge would have to be in the form of basic amino acids. Domain X of MATK possesses a much higher amount of basic amino acids than other regions of the ORF and contains the conserved α-helices αH and αI (Figs. 2A and 5). These features suggest that domain X in MATK most likely forms part of the template-primer binding tract as seen with LTRA (Blocker et al. 2005). Additional features found in the MATK ORF common to LTRA include the highly conserved α-helices αJ in domain X as well as insertions 3a and "ti," which were shown by unigenic evolution analysis as possibly important for maturase function (Cui et al. 2004; Blocker et al., 2005). The combination of these common structural and biochemical features between MATK and LTRA supports the proposed function of MATK as a group II intron maturase.

Previous analysis of sequence homology between MATK and LTRA identified only sequence blocks RT5, RT6, and RT7 to be conserved in the RT domain of MATK (Mohr et al. 1993). Our secondary structure alignment indicated that sequence blocks RT0 and RT3 are also retained as part of the RT domain in the MATK ORF (Fig. 5). The RT0 region has been suggested to be part of an extended fingers region required for successful binding of the RNA template (Cui et al. 2004). The RT0 block in LTRA is characterized by an α-helix in the N-terminal region containing a conserved alanine residue (Blocker et al. 2005). Both of these features are maintained in the RT0 region of MATK as determined by secondary structure alignment despite the almost-complete lack of sequence similarity (Fig. 5).

## MATK Nonpolar Composition in the Gnetophyta

The only significant deviation in side chain composition evident in MATK from across angiosperms was in the Gnetophyta. This group of gymnosperms had a 12.5 % increase in hydrophobic amino acids compared to other green plant taxa (Fig. 1B). Gnetophytes are an evolutionarily unique group of plants with morphological characteristics common to both gymnosperms and angiosperms (Carmichael and Friedman 1995; Donoghue and Doyle 2000) and a disputable position in land plant phylogeny (see Won and Renner 2006). The divergence in hydrophobic amino acid composition in MATK observed in the gnetophytes may be yet another unique trait to this enigmatic plant lineage. Further investigation is required to determine if the increase in MATK hydrophobicity is significant with reference to function of this protein in the gnetophytes.

## MATK Localization

The suborganelle location of MATK is currently unknown. However, identification of putative transmembrane helices in the MATK ORF suggests that this protein may be associated with a membrane. Similar transmembrane regions were found for the eukaryotic group II intron maturases, MATR, COB-I1, and COX1-I2. The lack of these transmembrane segments in LTRA is one of the few structural divergences between these two maturases. These results imply that the evolution of transmembrane regions, and, consequently, putative membrane association, in group II intron maturases is a trait acquired after the divergence of eukaryotic organisms from their prokaryotic ancestors.

## Evolutionary Constraint on MATK

Analysis of side chain composition in MATK across green plants displayed surprisingly little variation despite the accelerated nonsynonymous amino acid substitution rate in this protein (Fig. 1B). Further, comparison of the predicted MATK secondary structure from pine, rice, and Arabidopsis demonstrated remarkable homology in structure (Fig. 4). Thus, not only is the biochemical makeup of MATK conserved across a wide range of evolutionarily divergent plant lineages, but also its deduced secondary structure. This evolutionary constraint was evident across the entire MATK ORF. Previous studies based on nucleotide substitution rates have indicated that domain X, the chief domain required for maturase activity (Moran et al. 1994; Cui et al. 2004), is the most highly conserved region in MATK (Hilu and Liang 1997). Our analysis of variation in side chain composition, however, demonstrated relatively equal constraint in both domains (Fig. 2A), suggesting that regions of high functional and structural importance also reside in the N-terminal domain. Likewise, comparison of mutation rates in third versus first and second codon position in nonphotosynthetic plants demonstrated no significant difference in constraint on domain X versus the rest of the matK gene (Young and dePamphilis 2000). Division of the N-terminal region into sectors indicated that there were at least two conserved regions in this domain, sector 2 and sector 5 (Fig. 3A). Further molecular experimentation is required to determine the specific role of sector 2 in MATK function. Sector 5 overlaps a region with sequence similarity to sequence blocks 5, 6, and 7 of the RT domain of other group II intron maturases (Mohr et al. 1993). These RT sequence blocks have been shown in other maturases to function as part of the palm region (one of three structures that form the template-primer binding tract) required for efficient maturase activity (Cui et al. 2004). Detection of the RT domain (sector 5) and domain X (sector 6) as highly conserved regions by examination of SD in side chain composition supports the reliability of this method to identify regions of structural and functional importance in MATK.

High or low evolutionary constraints are relative terms and cannot be stated without comparison to proteins on either end of this spectrum. We compared variation in side chain composition of this protein to that of three other proteins, RBCL, INFA, and MATR in order to establish the mode and relative degree of evolutionary constraint on MATK. Variation in side chain composition of MATK was shown to be higher than that of the well-conserved chloroplast protein RBCL but not as fast as translated sequences from the pseudogene infA. Evolutionary constraint on MATK at this higher level of

protein structure, therefore, appears to fall somewhere in the middle of these two extremes. This placement would suggest that overall mutations in MATK are neutral and do not result in change (positive or negative) in protein structure. Müller et al. (2006), using a likelihood ratio test on the ratio of nonsynonymous to synonymous substitutions in angiosperm taxa, also found that *matK* evolves closer to neutrality. Further comparison of variation in MATK side chain composition to the mitochondrial maturase MATR revealed that the level of variation evident in MATK is typical of this kind of enzyme (Table 2).

## Conclusion

Evolutionary constraint is generally determined by the rate of nucleotide and amino acid substitution (Garcia-Maroto et al. 1991; Wolfe and dePamphilis 1998; Ophir et al. 1999; Halligan et al. 2004). Synonymous substitutions are regarded as neutral or silent mutations, whereas nonsynonymous substitutions are indicative of selective pressures (Ophir et al. 1999; Young and dePamphilis 2005). Conservative amino acid replacement, however, must be taken into account to obtain a more accurate view of evolution in rapidly evolving genes. By determining amino acid side chain composition and variation of this composition in the MATK ORF, we have shown that this rapidly evolving gene is under functional and structural constraint. Comparison of predicted secondary structure for this protein among three divergent taxa and to the LTRA group II intron maturase further supported our side chain composition data and demonstrated very strong structural constraint on MATK. This comparison of secondary structure has also revealed that MATK contains several structural features required for maturase function, including the "ti" insertion inherent to the LTRA group II intron maturase but not the more distantly related HIV-1 RT (Blocker et al. 2005). Taken together, these results have uncovered structural and functional information for MATK that can aid in molecular investigations of this important maturase, including the identification of a new region (sector 2) that may contribute to MATK maturase activity. We have also demonstrated the utility of using side chain composition and variation in this composition in determining evolutionary constraint in rapidly evolving genes.

## References

Albert VA, Backlund A, Bremer K, Chase MW, Manhart JR, Mishler BD, Nixon KC (1994) Functional constraints and rbcL evidence for land plant phylogeny. Ann Mo Bot Garden 81:534–567

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The shape and structure of proteins. In: Molecular biology of the cell. Garland Sciences, New York, Fig. 3.1

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. AAAI Press, Menlo Park, CA

Barthet MM, Hilu KW (2007) Expression of *matK*: functional and evolutionary implications. Am J Bot 94:1402–1412

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) The Pfam protein families database. Nucleic Acids Re 30:276–280

Blocker FJH, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM (2005) Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11:14–28

Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89:2002–2006

Cameron KM (2005) Leave it to the leaves: a molecular phylogenetic study of Malaxideae (Orchidaceae). Am J Bot 92:1025–1032

Carmichael JS, Friedman WE (1995) Double fertilization in *Gnetum gnemon*: the relationship between the cell cycle and sexual reproduction. Plant Cell 7:1975–1988

Chase MW, Soltis DE, Olmstead RG, 42 coauthors (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann Mo Bot Gardens 80:528–580

Cuff JA, Barton GJ (2000) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40:502–511

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Jpred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM (2004) A group II intron-encoded maturase functions preferentially *in cis* and requires both the reverse transcriptase and X domains to promote RNA splicing. J Mol Biol 340:211–231

Donoghue MJ, Doyle JA (2000) Seed plant phylogeny: Demise of the anthophyte hypothesis? Curr Biol 10:R106–R109

Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, Gerstein M (2002) Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 30:2515–2523

Ems S. C, Morden CW, Dixon CK, Wolfe KH, dePamphilis CW, Palmer JD (1995) Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. Plant Mol Biol 29:721–733

Farré J, Araya A (1999) The *mat-r* open reading fram is transcribed from a non-canonical promoter and contains an internal promoter to co-transcribe exons *nad1e* and *nad5III* in wheat mitochondria. Plant Mol Biol 40:959–967

Garcia-Maroto F, Castagnaro A, de la Hoz P, Maraña C, Carbonero P, García-Olmedo F (1991) Extreme variations in the ratios of nonsynonymous to synonymous nucleotide substitution rates in signal peptide evolution. FEBS Lett 287:67–70

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Graur D, Li W-H (1988) Evolution of protein inhibitors of serine proteinases: Positive Darwinian selection or compositional effects?. J Mol Evol 28:131–135

Halligan DL, Eyre-Walker A, Adolfatto P, Keightley PD (2004) Patterns of evolutionary constraints in intronic and intergenic DNA of Drosophila. Genome Res 14:273–279

Hayashi K, Kawano S (2000) Molecular systematics of *Lilium* and allied genera (Liliaceae): phylogenetic relationships among *Lilium* and related genera based on the *rbcL* and *matK* gene sequence data. Plant Species Biol 15:73–93

Hilu KW, Liang H (1997) The *mat*K gene: sequence variation and application in plant systematics. Am J Bot 84:830–839

Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Jens GR, Campbell CS, Chatrou LW (2003) Angiosperm phylogeny based on *matK* sequence information. Am J Bot 90:1758–1776

Jenkins BD, Khulhanek DJ, Barkan A (1997) Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classes with distinct requirements for splicing factors. Plant Cell 9:283–296

Johnson LA, Soltis DE (1994) MatK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Syst Bot 19:143–156

Jukes TH, Kimura M (1984) Evolutionary constraints and the neutral theory. J Mol Evol 21:90–92

Kellogg EA, Juliano ND (1997) The structure and function of RuBisCO and their implications for systematic studies. Am J Bot 84:413–428

Lodish H, Berk A, Zipursky LS, Matsudaira P, Baltimore D, Darnell J (2000) Molecular cell biology. W. H. Freeman, New York

Magallón S, Sanderson MJ (2002) Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. Am J Bot 89:1991–2006

Matsuura M, Saldanha R, Ma H, Wank H, Yang, Mohr G, Cavanagh S, Dunny GM, Belfort M, Labmowitz AM (1997) A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. Genes Dev 11:2910–2924

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13: 645–658

Mohr G, Perlman PS, Lambowitz AM (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Res 21:4991–4997

Moran JV, Mecklenburg, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman PS (1994) Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Res 22:2057–2064

Müller KF, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK, trnT-F* and *rbcL* in basal angiosperms. Mol Phylogenet Evol 41:99–117

Neuhaus H, Link G (1987) The chloroplast tRNA$^{Lys}$ (UUU) gene from mustard (*Sinapsis alba*) contains a class II intron potentially coding for a maturase-related polypeptide. Curr Genet 11:251–257

Olmstead RG, Palmer JD (1994) Chloroplast DNA systematics: a review of methods and data analysis. Am J Bot 81:1205–1224

Ophir R, Itoh T, Graur D, Gojobori T (1999) A simple method for estimating the intensity of purifying selection in protein-coding genes. Mol Biol Evol 16:49–53

Persson B, Argos P (1994) Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. J Mol Biol 237:182–192

Rambo RP, Doudna JA (2004) Assembly of an active group II intron-maturase comlex by protein dimerization. Biochemistry 43:6486–6497

Rost B (2001) Review: Protein secondary structure prediction continues to rise. J Struct Biol 134:204–218

Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232:584–599

Rost B, Yachdav G, Liu J (2004) The PredictProtein Server. Nucleic Acids Res 32 (Web Server issue):W321–W326

Saldanha R, Chen B, Wank H, Matsuura M, Edwards J, Lambowitz AM (1999) RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38:9069–9083

Sander C, Schneider R (1991) Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 9:56–68

Soltis DE, Soltis PS (1998) Choosing an approach and an appropriate gene for phylogenetic analysis. Kluwer Academic, Boston

Swofford DL (2001) PAUP: Phylogenetic Analysis Using Parsimony, version 4.0b6. Sinauer, Sunderland, MA

Tourasse NJ, Li W-H (2000) Selective constraints, amino acid composition, and the rate of protein evolution. Mol Biol Evol 17:656–664

Vogel J, Hubschmann T, Borner T, Hess WR (1997) Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for MatK as an essential splicing factor. J Mol Biol 270:179–187

Vogel J, Borner T, Hess W (1999) Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Res 27:3866–3874

Wolfe AD, dePamphilis CW (1998) The effect of relaxed functional constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. Mol Biol Evol 15:1243–1258

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natal Acad Sci USA 84:9054–9058

Won H, Renner SS (2006) Dating dispersal and radiation in the gymnosperm *Gnetum* (Gnetales)—Clock calibration when outgroup relationships are uncertain. Syst Biol 55:610–622

Xia X, Li W-H (1998) What amino acid properties affect protein evolution?. J Mol Evol 47:557–564

Xiang Q-Y, Soltis DE, Morgan DR, Soltis PS (1998) Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences. Am J Bot 85:285–297

Young ND, dePamphilis CW (2000) Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Mol Biol Evol 17:1933–1941

Young ND, dePamphilis CW (2005) Rate variation in parasitic plants: correlated and uncorrelated patterns among plastid genes of different function. BMC Evol Biol 5:16–25

Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 195:957–961