Amino Acids
DOI 10.1007/s00726-011-0872-8

ORIGINAL ARTICLE

Predicting sub-cellular localization of tRNA synthetases from their primary structures

Bharat Panwar · G. P. S. Raghava

Received: 11 October 2010 / Accepted: 21 February 2011

Abstract Since endo-symbiotic events occur, all genes of mitochondrial aminoacyl tRNA synthetase (AARS) were lost or transferred from ancestral mitochondrial genome into the nucleus. The canonical pattern is that both cytosolic and mitochondrial AARSs coexist in the nuclear genome. In the present scenario all mitochondrial AARSs are nucleus-encoded, synthesized on cytosolic ribosomes and post-translationally imported from the cytosol into the mitochondria in eukaryotic cell. The site-based discrimination between similar types of enzymes is very challenging because they have almost same physico-chemical properties. It is very important to predict the sub-cellular location of AARSs, to understand the mitochondrial protein synthesis. We have analyzed and optimized the distinguishable patterns between cytosolic and mitochondrial AARSs. Firstly, support vector machines (SVM)-based modules have been developed using amino acid and dipeptide compositions and achieved Mathews correlation coefficient (MCC) of 0.82 and 0.73, respectively. Secondly, we have developed SVM modules using position-specific scoring matrix and achieved the maximum MCC of 0.78. Thirdly, we developed SVM modules using N-terminal, intermediate residues, C-terminal and split amino acid composition (SAAC) and achieved MCC of 0.82, 0.70, 0.39 and 0.86, respectively. Finally, a SVM module was developed using selected attributes of split amino acid composition (SA-SAAC) approach and achieved MCC of 0.92 with an accuracy of 96.00%. All modules were trained and tested on a non-redundant data set and evaluated using fivefold cross-validation technique. On the independent data sets, SA-SAAC based prediction model achieved MCC of 0.95 with an accuracy of 97.77%. The web-server '*MARSpred*' based on above study is available at http://www.imtech.res.in/raghava/marspred/.

Keywords Mitochondrial tRNA synthetase · Support vector machine · Prediction · MARSpred

B. Panwar · G. P. S. Raghava (✉)
Bioinformatics Centre, Institute of Microbial Technology (CSIR), Sector 39A, Chandigarh, India
e-mail: raghava@imtech.res.in
URL: http://www.imtech.res.in/raghava/

Introduction

In the process of evolution, most of the bacterial genes present in the ancestral organellar genomes have been either disappeared or transferred to the nucleus (Doolittle 1998). In this process, all genes of mitochondrial AARSs were also lost or shifted into the nucleus. The eukaryotic nucleus genome codes two different sets of AARS for cytosol and mitochondria. These mitochondrial AARSs are post-translationally imported into the mitochondria (Brindefalk et al. 2007). The function of AARSs is to precisely attach correct amino acids with tRNAs containing the corresponding anticodon (Berg 1961). There are twenty AARSs found in maximum number of organisms and mainly each one is specific for single amino acid (Rajbhandary 1997). The translation of few but crucial protein-encoding genes, remaining on the mitochondrial genome (Unseld et al. 1997), requires complete set of mitochondrial tRNA synthetases in mitochondria where translation occurs. The imported protein is guided through the import complexes by a targeting sequence at the N-terminal part of the protein (Baker et al. 2007). A study reported that both cytosolic and mitochondrial tRNA synthetases are essential for cell survival and are not interchangeable in *T. brucei* (Español et al. 2009). The defected AARSs can be lethal and lead to

Published online: 13 March 2011

numerous pathological problems including cancer, neuronal pathologies, autoimmune disorders, and disrupted metabolic conditions (Antonellis and Green 2008; Schimmel 2008; Park et al. 2008; Lee et al. 2006). The mutant gene of mitochondrial aspartyl-tRNA synthetase causes leukoencephalopathy, which has brain-stem and spinal cord involvement and lactate elevation (LBSL) (Scheper et al. 2007). The mutated promoter region of mitochondrial isoleucyl-tRNA synthetase modifies its expression in hereditary non-polyposis colorectal cancer (HNPCC) and Turcot syndrome (Miyaki et al. 2001). A single nucleotide polymorphism of mitochondrial leucyl-tRNA synthetase leading to an amino acid substitution (H324Q) was found in patients afflicted with type 2 diabetes mellitus (t Hart et al. 2005). The functional annotation tools for mitochondrial-tRNA synthetases can help biologists to better understand these diseases.

The sub-cellular location of AARSs can be predicted by machine learning techniques. Earlier many computational methods have been developed for the prediction of mitochondrial proteins (Guda et al. 2004; Kumar et al. 2006). But these are comprehensive method and act universally for all diverse type of mitochondrial proteins. To investigate this problem, we have developed a tool for the discrimination between cytosolic and mitochondrial tRNA synthetases using support vector machine (SVM). In this process, we analyzed both the type of tRNA synthetases and optimized maximum distinguishable features from protein sequences. First we applied amino acid composition; dipeptide composition; PSSM and split amino acid composition (SAAC) patterns-based approaches. We have developed a prediction tool '*MARSpred*' using selected composition of 40 residues of N-terminal, 60 residues of C-terminal and intermediate amino acids. The SA-SAAC based prediction model also performed well on independent data sets. This prediction tool will be very helpful for the functional annotation of tRNA synthetases by discrimination between cytosolic and mitochondrial AARSs.

Results

It is very difficult to find out distinctive patterns from the similar type of enzymes, which are only different in their subcellular location. The machine learning for the prediction tools development requires distinguishable features. We have used the 40% non-redundant sequences of both mitochondrial and cytosolic AARSs. The composition of physico-chemical properties of both cytosolic and mitochondrial AARSs was calculated. It was observed that they have almost equal properties (Fig. 1). We have used many distinguishable patterns for SVM-based machine learning. Different kernels and parameters of SVM were tried and optimized the best performance for discrimination between mitochondrial (positive) and cytosolic (negative) tRNA synthetases.

Sequence similarity search

One of the common practices for functional annotation of a new protein is to perform a sequence similarity search

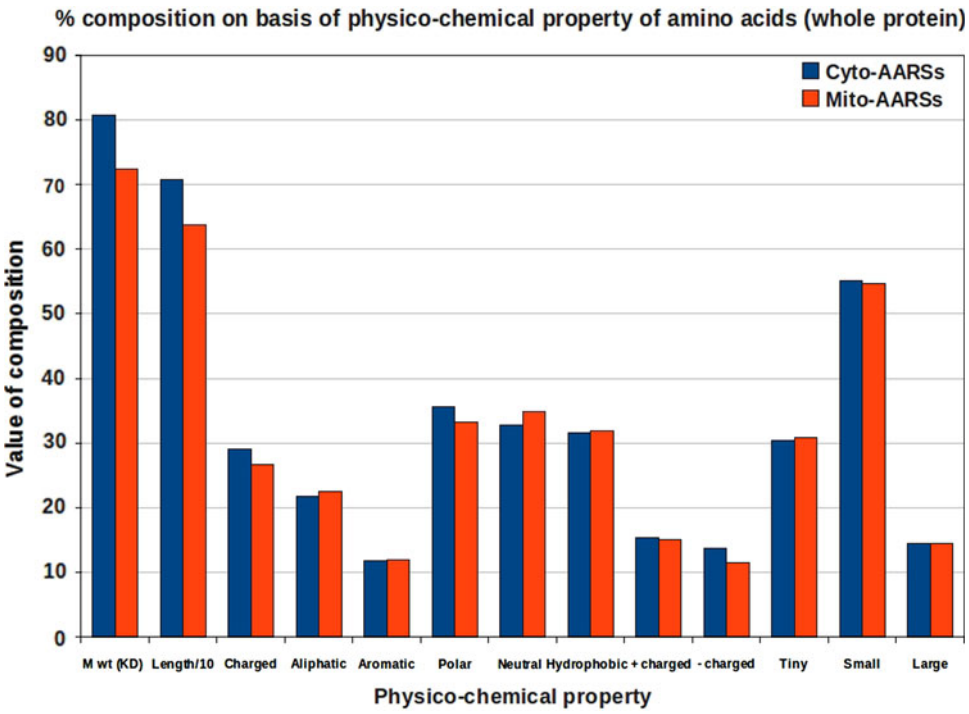

Fig. 1 The composition of physico-chemical properties of cytosolic and mitochondrial tRNA synthetases

against a database of well-annotated proteins. Thus, we developed BLAST-based approach for discriminating cytosolic and mitochondrial tRNA synthetases. At the 1 *E* value threshold, all query sequences were found their target sequence (see detail in "Materials and methods"). We have achieved maximum 81.36% sensitivity, 70.73% specificity, 77.00% accuracy with 0.52 MCC. This demonstrates that BLAST alone can discriminate only 81.36% mitochondrial tRNA synthetases from cytosolic tRNA synthetases. Thus, there is a need to develop prediction models based on machine learning techniques. We applied various amino acid composition, dipeptide composition and position-specific scoring matrix (PSSM)-based approaches to discriminate cytosolic and mitochondrial tRNA synthetases with high accuracy.

Amino acid compositions-based approach

It has been shown in the past that amino acid composition can be used to classify the diverse class of proteins and development of prediction tools using machine learning techniques (Raghava and Han 2005; Garg et al. 2005). We calculated the amino acid composition of both type of tRNA synthetases and observed that they have significant difference from each other (Fig. 2). SVM-based classifier was developed using 20 dimension vectors of amino acid composition, one for each amino acid. We achieved 98.33% sensitivity, 80.28% specificity, 91.00% accuracy and 0.82 MCC.

Dipeptide composition-based approach

In the previous studies, it has been shown that dipeptide composition-based methods are more successful than amino acid-composition based in the discrimination between different class of proteins (Bhasin and Raghava 2004). We calculated dipeptide composition of both type of tRNA synthetases. SVM-based classifier was developed using 400 dimensions of vector (20 × 20) of dipeptide compositions, one for each dipeptide. We achieved 83.18% sensitivity, 90.00% specificity, 86.00% accuracy and 0.73 MCC.

PSSM-based approach

In the past, multiple sequence alignment information in form of PSSM has been used for developing prediction methods (Kaur and Raghava 2004; Kumar et al. 2007). First we created PSSM profile for each protein using position-specific iterative BLAST (PSI-BLAST) search against Swiss-Prot database. Secondly, we computed a vector of dimension of 400 (20 × 20) from PSSM matrix. Finally a SVM model was developed using PSSM and achieved 88.33% sensitivity, 90.00% specificity, 89.00% accuracy with 0.78 MCC.

N-terminal residues-based approach

In the mitochondrial tRNA synthetases, N-terminal residues are responsible for the targeting of tRNA synthetases

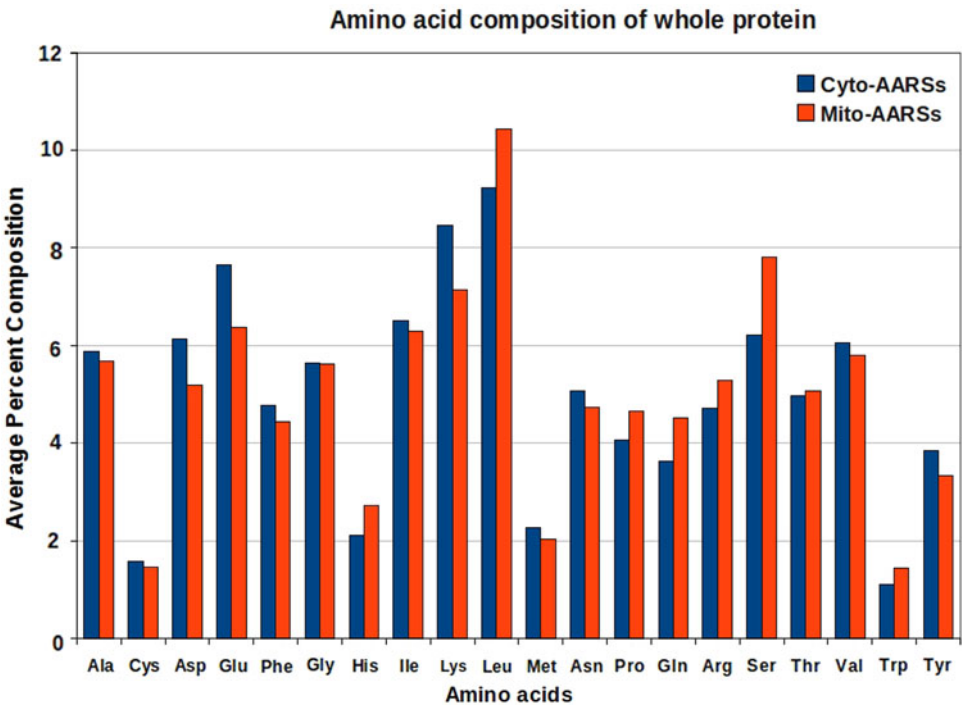

Fig. 2 The amino acid composition of cytosolic and mitochondrial tRNA synthetases

into mitochondrial tRNA synthetases (Duchêne et al. 2009). So we have calculated different N-terminal amino acid compositions of cytosolic and mitochondrial AARSs. We compared various length of N-terminal and observed that amino acid composition of 40 residues of mitochondrial N-termini is significantly different from cytosolic tRNA synthetases (Fig. 3). The SVM module achieved 95.00% sensitivity, 85.00% specificity, 91.00% accuracy and 0.82 MCC.

C-terminal residues-based approach

Recently, one study suggested that C-terminal is responsible for the activity of tRNA synthetases (Español et al. 2009). We calculated different C-termini amino acid compositions and observed that composition of 60 residues of C-terminal is different in mitochondrial and cytosolic AARSs (Fig. 4). The SVM module achieved maximum 59.85% sensitivity, 78.33% specificity, 67.00% accuracy and 0.39 MCC.

Intermediate residues approach

After amino acid composition calculation of N-terminal (40 residues) and C-terminal (60-residues), we calculated composition of remaining intermediate amino acids (Fig. 5). The best parameter of SVM-based machine learning was optimized and achieved 79.85% sensitivity, 90.00% specificity, 84.00% accuracy and 0.70 MCC.

SAAC-based approach

We calculated SAACs for 40 amino acids of N-terminus, 60 amino acids of C-terminus and intermediate amino acids. We have developed SVM module of all these compositions and achieved 93.18% sensitivity, 92.50% specificity, 93.00% accuracy with 0.86 MCC.

Selected attributes of split amino acid composition (SA-SAAC) based approach

In this approach, we selected significant attributes (amino acids) using WEKA 3.6.0 version. We have selected a total of 13 attributes, which were 4, 1 and 8 from N-termini, C-termini and intermediate regions, respectively. All protein sequences were divided into three parts (N-40, C-60 and intermediate) and total of 13 split amino acids compositions of selected amino acids were calculated (Fig. 6). We developed highly efficient SVM module and achieved 98.33% sensitivity, 92.50% specificity, 96.00% accuracy and 0.92 MCC. In all cases we found that SA-SAAC based approach performed better than others. The SA-SAAC based prediction model achieved 100% sensitivity, 96.69% specificity, 97.77% accuracy and 0.95 MCC on the independent data sets (Table 1). These results confirmed that performance of SA-SAAC based model is not biased for our 40% non-redundant main data set only.

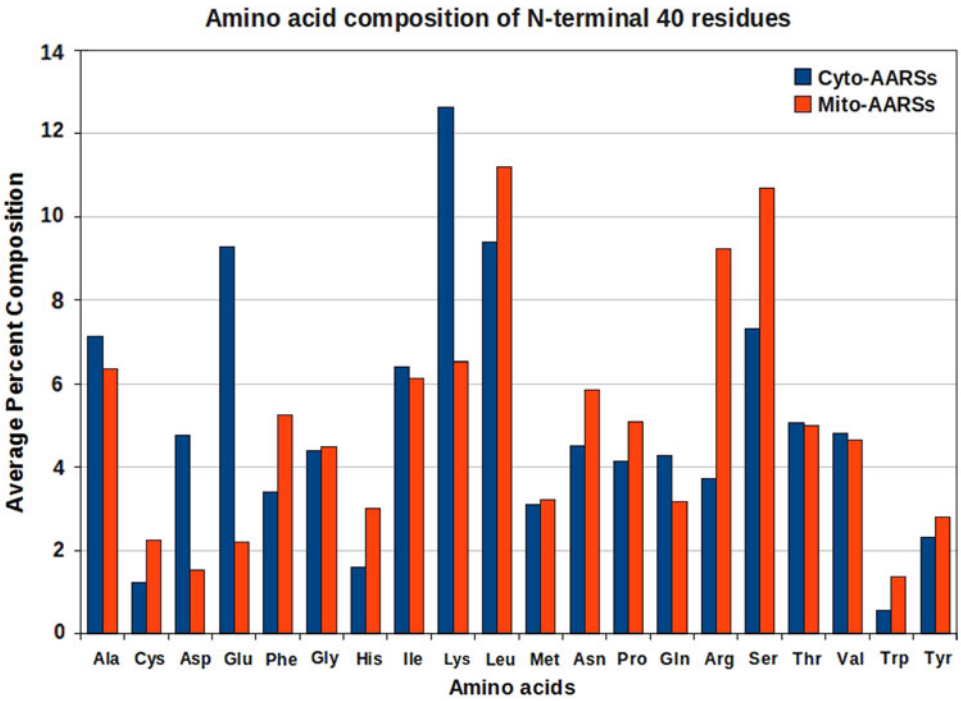

Fig. 3 The N-termini (40 residues) amino acid composition of cytosolic and mitochondrial tRNA synthetases

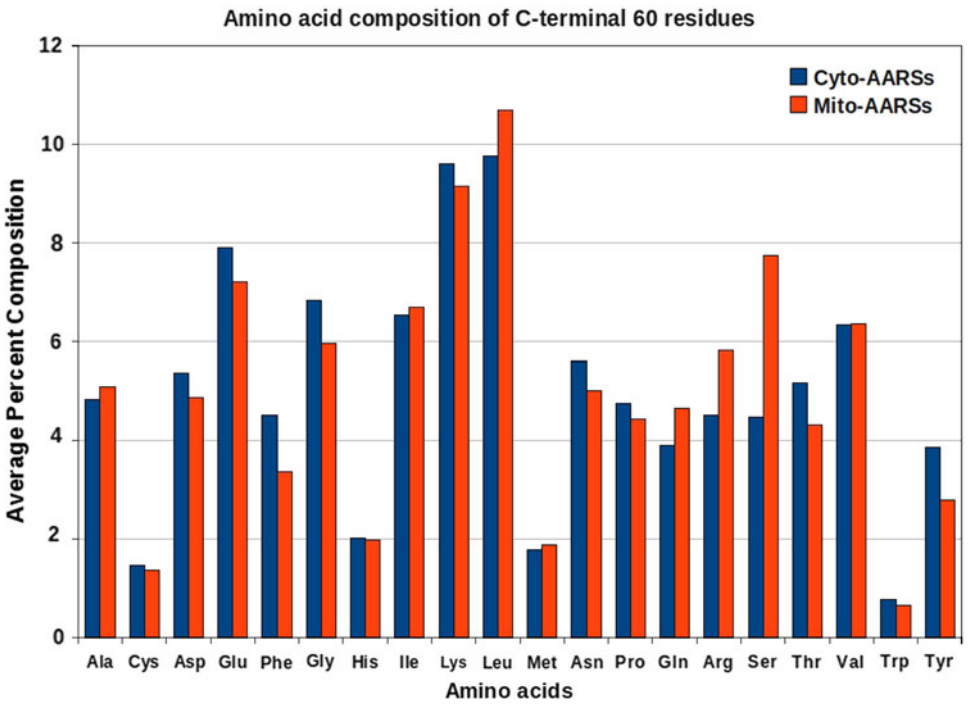

Fig. 4 The C-termini (60 residues) amino acid composition of cytosolic and mitochondrial tRNA synthetases

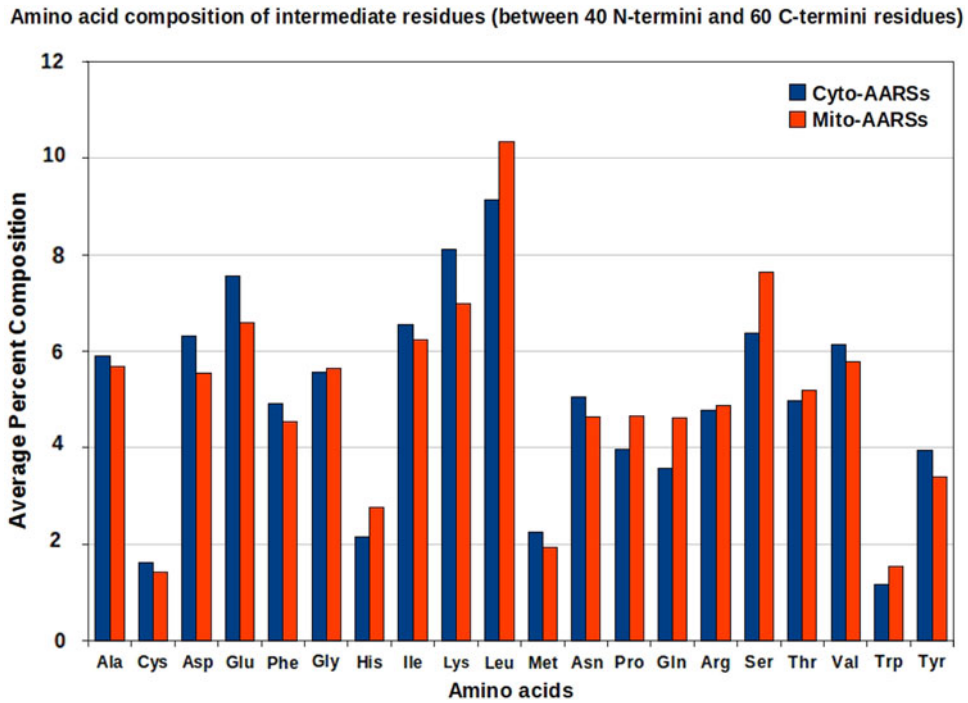

Fig. 5 The intermediate (between 40 residues of N-termini and 60 residues of C-termini) amino acid composition of cytosolic and mitochondrial tRNA synthetases

Performance comparisons with other algorithms

We have compared other external softwares with our SA-SAAC based method (MARSpred). In this comparison, we have used independent data sets, which were not used for the training of our prediction method. The MARSpred, MitoProt (Claros and Vincens 1996), MitPred (Kumar et al. 2006), MITOPRED (Guda et al. 2004) and Predotar (Small et al. 2004) algorithms achieved MCC of 0.95, 0.87, 0.81, 0.77 and 0.76, respectively (Table 2). All these methods have been provided with different parameters. We have used all available parameters and find out maximum possible performance of these methods. These result shows that SA-SAAC based approach is more robust and efficient for the discrimination between mitochondrial and cytosolic tRNA synthetases.

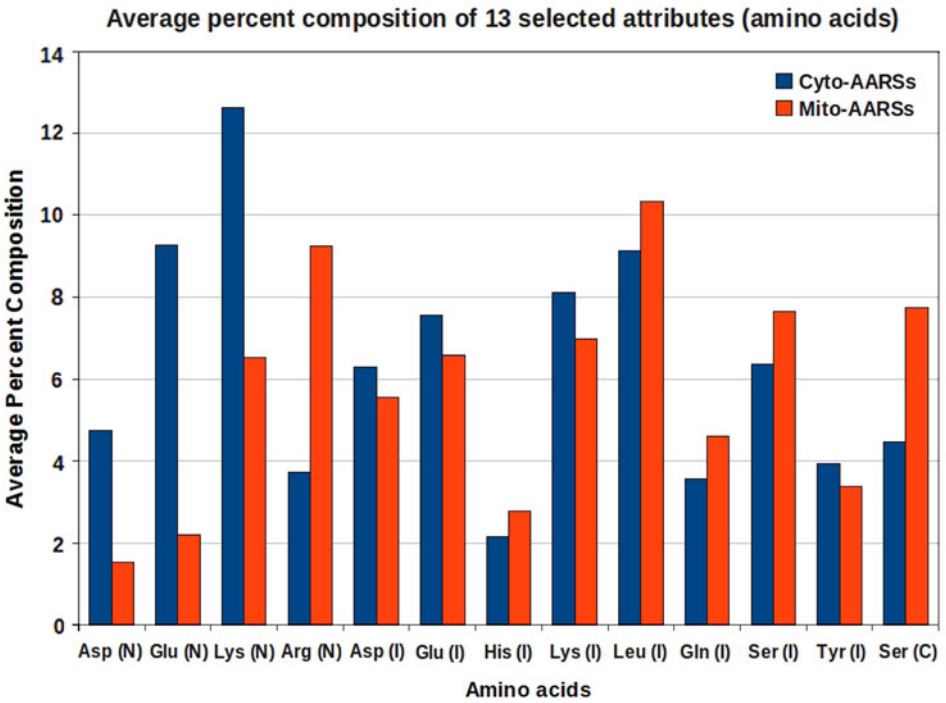

Fig. 6 The average amino acid composition of 13 selected attributes for SA-SAAC appraoch

Table 1 The performance of SVM for different approaches

S. no.	Approach	Sensitivity	Specificity	Accuracy	MCC
1	BLAST-based approach (E value = 1)	81.36	70.73	77.00	0.52
2	Amino acid composition based	98.33	80.28	91.00	0.82
3	Dipeptide composition based	83.18	90.00	86.00	0.73
4	PSSM based	88.33	90.00	89.00	0.78
5	N-terminal (40 residues) based	95.00	85.00	91.00	0.82
6	C-terminal (60-residues) based	59.85	78.33	67.00	0.39
7	Intermediate residues based	79.85	90.00	84.00	0.70
8	SAAC approach	93.18	92.50	93.00	0.86
9	SA-SAAC approach	**98.33**	**92.50**	**96.00**	**0.92**
10	SA-SAAC approach (independent datasets)	**100.00**	**96.69**	**97.77**	**0.95**

Bold values indicate high performances

Table 2 The performance comparisons of MARSpred with other algorithms

S. no.	Approach	Sensitivity	Specificity	Accuracy	MCC	Parameters
1	MARSpred	**100.00**	**96.69**	**97.77**	**0.95**	Threshold = **0.7**
2	MitoProt	96.55	92.56	93.85	0.87	Probability = 0.4
3	MitPred	81.03	96.69	91.62	0.81	Threshold = −0.5
4	MITOPRED	94.83	85.95	88.83	0.77	Confidence cut-off 60%
5	Predotar	68.97	99.17	89.39	0.76	Default

All the methods are ordered by MCC value

Bold values indicate high performances

Functional annotation of AARSs

We have created a data set of total 88 unknown AARSs. These are experimentally validated eukaryotic AARSs but the sub-cellular localization of sequences are still unclear. It is very important to predict their sub-cellular localization to understand their role in protein synthesis properly. The SA-SAAC based method predicted 17 and 71 protein sequences as mitochondrial and cytosolic AARSs, respectively.

It is interesting to create organism-specific whole sets of mitochondrial and cytosolic-tRNA synthesis but the availability of experimentally validated AARSs is very low. Thus, we have retrieved total of 5,557 automatically annotated eukaryotic AARSs from TrEMBL. These AARSs requires further location-based functional annotation. The SA-SAAC based method predicted 1,976 and 3,581 protein sequences as mitochondrial and cytosolic AARSs, respectively. We have prepared different sets for *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. In the sets making process, first we have selected all well annotated and location-wise experimentally validated AARSs. Secondly, we predicted remaining (location information is not known) experimentally validated AARSs (marks with superscript "a") with our SA-SAAC based approach. Thirdly, we have predicted automatically annotated AARSs (marks with superscript "b") from TrEMBL and tried to make whole sets for each organism. We have found whole sets in *Homo sapiens* only, other organisms lack either mitochondrial or cytosolic amino acid-specific AARSs. Table 3 contains UniProt ID of organism-wise available sets for mitochondrial and cytosolic AARSs. We have discriminated these protein sequences of AARSs based on the predicted SVM scores. The positive (>0) and negative (<0) scores predicted as mitochondrial and cytosolic tRNA synthetases, respectively.

Discussion

All nucleus-encoded AARSs are first resides into cytosol. Some of these AARSs transported from cytosol into mitochondria. The experimental determination of sub-cellular location of mitochondrial AARSs are very labour-intensive and time-consuming procedure. To assist the biologists in assigning the function of unknown AARSs protein, a systematic attempt has been made for predicting the site of AARSs. We obtained both mitochondrial (positive) and cytosolic (negative) protein sequences from UniProt database. We have selected many distinguishable features between mitochondrial and cytosolic AARSs. Amino acid composition-based comparison study suggested that mitochondrial and cytosolic AARSs prefer different amino acids for their construction.

The comparative analysis of mitochondrial and cytosolic tRNA synthetases revealed that 40 residues of N-terminal and 60 residues of C-terminal were significantly different (Figs. 3, 4). We applied different type of approaches (PSSM, amino acid, dipeptide, N-terminal, C-terminal, SAAC and SA-SAAC) in machine learning of SVM. We achieved maximum SVM performance in SA-SAAC based approach. We found that Leu and Ser are more abundant in mitochondrial tRNA synthetases and Asp, Glu and Lys are preferred in cytosolic tRNA synthetase.

It is well established that N-terminal contains signal peptide for sub-cellular localization of mitochondrial proteins. In the N-terminal region (40 residues) Leu, Arg and Ser are more profuse in mitochondrial tRNA synthetases and Asp, Glu and Lys are favoured in cytosolic tRNA synthetase. In the case of C-terminal region (60 residues) Leu, Gln, Arg and Ser are more abundant in mitochondrial tRNA synthetases and Glu, Phe, Gly, Lys and Tyr are preferred in cytosolic tRNA synthetase. In the intermediated region between N- and C-terminal Leu, Pro, Gln and Ser are more plentiful in mitochondrial tRNA synthetases and Asp, Glu and Lys are favoured in cytosolic tRNA synthetases.

We have selected 13 most distinguishable features between mitochondrial and cytosolic tRNA synthetases using Weka. These were four (Asp, Glu, Lys, Arg), one (Ser) and eight (Asp, Glu, His, Lys, Leu, Gln, Ser, Tyr) amino acids from N-termini, C-termini and intermediate regions, respectively (Fig. 6). The SA-SAAC based method performed well on both main and independent data sets. We have predicted different sets of mitochondrial and cytosolic AARSs for *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *A. thaliana* and *S. cerevisiae*. In future, experimental validation of mitochondrial and cytosolic AARSs is very essential because their availability is very low now. It will be very interesting to know how eukaryotic cell maintains and regulates these two different types of essential enzyme of translation machinery. We anticipate that our prediction method '*MARSpred*' will help researchers to better understand the symbiotic affiliation between mitochondria and eukaryotic cell.

Conclusions

To conclude, the present work is an attempt to discriminate AARSs on the basis of their sub-cellular location. We analyzed protein sequences of both mitochondrial and cytosolic AARSs and selected the distinguishable patterns. These were amino acid, dipeptide, PSSM, N-terminal, C-terminal and SAACs. We used these features as a SVM input-based machine learning. We were able to model an efficient classifier from selected SAAC-based information. A server called *MARSpred* was developed on the results obtained.

Materials and methods

Data sets

The eukaryotic AARSs data of total 390 experimentally validated protein sequences were obtained from the

Table 3 The organism-specific predicted sets for mitochondrial and cytosolic AARSs using SA-SAAC approach

AARSs	Homo sapiens		Mus musculus		Drosophila melanogaster		Caenorhabditis elegans		Arabidopsis thaliana		Saccharomyces cerevisiae	
	Cyto	Mito	Cyto	Mito	Cyto	Mito	Cyto	Mito	Cyto	Mito	Cyto	Mito
AlaRS	P49588	Q5JTZ9	Q8BGQ7	Q14CH7	Q9VLM8	Q9VRJ1	O01541	Q23122	P36428-2	P36428-1	P40825	N/A
ArgRS	P54136	Q5T160	Q9D0I9	Q3U186	Q9VXN4	Q8SXK2[b]	Q19825	Q18316[b]	Q9C713[b]	O23247[b]	Q05506	P38714
AsnRS	O43776	Q96I59	Q8BP47	Q8BGV0	Q9V434[b]	Q7JZR5[b]	Q19722	N/A	Q9SW96 Q9SW95 Q9SSK1	O48593	P38707	P25345
AspRS	P14868	Q6PI48	Q922B2	Q8BIP0	Q7K0E6[b]	Q9VJH2[b]	Q03577	P90831[b]	Q9SZ27[b]	O81892[b]	P04802	P15179
CysRS	P49589	Q9HA77	Q9ER72	Q8BYM8	Q7KN90	N/A	N/A	O76618[b] (a) Q965I6[b] (b)	Q0WLQ1[b]	O82267[b]	N/A	P53852[a]
GlnRS	P47897[a]	B4DTH6[b]	Q8BU21[b] Q8BML9[b]	N/A	Q9Y105[a]	N/A	O62431[a]	O44413[b] (bi)	A4UVN3[b]	N/A	P13188[a]	N/A
GluRS	P07814 (bi)	Q5JPH6	Q8CGC7 (bi)	Q9CXJ1	P28668 (bi)	Q9VV59[b]	Q23315[b]	O44413[b] (bi)	O82462[b]	Q9FEA2	P46655	P48525
GlyRS	P41250	P41250	N/A	Q9CZD3[a]	Q961R8[b]	C9QP25[b]	Q8I711[b]	Q10039[a]	Q9FXG2	O23627	P38088[a]	Q06817[a]
HisRS	P12081	P49590	Q61035	Q99KK9	Q9VWT1[b]	Q9VUK8[b]	N/A	P34183[a]	Q9M8R8[b]	O82413[b]	N/A	P07263
IleRS	P41252	Q9NSE4	Q8BU30	Q8BIJ6	Q8MSW0[b]	Q9VUY4[b]	Q21926	Q7Z261[b]	Q9SV89[b]	Q8RXK8[b]	P09436	P48526
LeuRS	Q9P2J5	Q15031	Q8BMJ2	Q8VDC0	Q9VQR8[b]	Q9VZ82[b]	Q09996[a]	Q23511[b]	Q56WB9[b]	Q9XEA0[b]	P26637	P11325
LysRS	Q15046[a]	Q15046-2	Q99MN1[a]	Q8CFK5[b]	Q8SXM8[b]	Q9W327[b]	Q22099[a]	Q56ZE1[b]	Q9ZPI1[a]	Q9LJE2[b]	P15180	P32048
PheRS	Q9Y285 (alpha) Q9NSD9 (beta)	O95363	Q8C0C7 (alpha) Q9WUA2 (beta)	Q99M01	Q9W3J5 (alpha) Q9VCA5 (beta)	O16129	D6VPB3[b] Q19713 (beta)	Q9U1Z3[b]	Q9T034 (alpha) Q9SGE9 (beta)	Q94K73	P15625 (alpha) P15624 (beta)	P08425
MetRS	P56192	Q96GW9	Q68FL6	Q499X9	N/A	Q9VFL5	Q20970	N/A	Q9SVN5[a]	Q9M2T9[b]	P00958	P22438
ProRS	P07814 (bi)	Q7L3T8	Q8CGC7 (bi)	Q8CFI5	P28668 (bi)	Q9VZY9[b]	Q22620[b] (a) Q65ZK5[b] (b)	O45869[b]	N/A	Q9FYR6[b]	P38708[a]	P39965
SerRS	P49591	Q9NP81	P26638	Q9JJL8	Q9VQL1[b]	Q9VF85[b]	Q18678	O45887[b]	Q39230[a]	Q8RWT8[b]	P07284	P38705
ThrRS	P26639 A2RTX5	Q9BW92	Q9D0R2	Q3UQ84	Q8IP94[b]	Q9VKB0[b]	P52709	N/A	Q8GZ45	O04630	P04801	P07236
TrpRS	P23381	Q9UGM6	P32921	Q9CYK1	Q0KI98[b]	Q9VVL8[b]	Q9U1R2[b]	P46579	Q9SR15[b]	Q8RXE9[b]	Q12109	P04803
TyrRS	P54577	Q9Y2Z4	Q91WQ3	Q8BYL4	Q9VV60[b]	Q9W107	Q8WQA5[b]	Q94262[b]	P93018[b]	Q9M876[b]	P36421	P48527
ValRS	P26640	Q5ST30	N/A	Q3U2A8	Q0E993[b]	Q9VSR7[b]	Q9U1Q4[a]	Q23360[b]	Q56X13[b]	P93736[a]	N/A	P07806

The UniProt ID shows for each AARS

Bi bifunctional tRNA synthetase

[a] Experimentally validated AARSs but location information is not known

[b] Automatically annotated AARSs from TrEMBL

UniProt database. This data set contains 162 cytosolic and 117 mitochondrial tRNA synthetases. The remaining 111 sequences containing 88 unknown AARSs (localization non-specified in the UniProt database), 7 chloroplastic AARSs and 16 AARSs are fragments or dual localized (mitochondrial/chloroplastic). We removed the sequence similarity of mitochondrial and cytosolic tRNA synthetases from the CD-HIT software (Li and Godzik 2006) and created a 40% non-redundant data set, which contained protein sequences of total 41 cytosolic and 59 mitochondrial tRNA synthetases. These 40% non-redundant data sets were used as main data sets. The remaining 58 mitochondrial and 121 cytosolic-tRNA synthetases were used as independent data sets. We have used 88 unknown AARSs (unknown dataset) for the location-based functional annotation of AARSs. We have also retrieved total of 5,557 automatically annotated eukaryotic AARSs from TrEMBL (dated 30-12-2010), which were used to make the organism-specific whole sets for mitochondrial and cytosolic AARSs. We have used mitochondrial tRNA synthetases as positive and cytosolic tRNA synthetases as negative data set for the development of the tools for discrimination between these two types of synthetases.

BLAST based approach

In this study we used BLAST for discriminating cytosolic and mitochondrial tRNA synthetases using fivefold cross-validation, where four sets of mitochondrial and cytosolic tRNA synthetases were used to create a BLAST database and both type of tRNA synthetases of the corresponding test set were searched against this BLAST database. This process was repeated five times, so the BLAST search was performed once for each mitochondrial and cytosolic tRNA synthetase. At 1 *E* value threshold all sequences of test sets found their target in BLAST database. We calculated the performance of BLAST in terms of sensitivity, specificity, accuracy and MCC.

Amino acid and dipeptide composition

The aim of calculating the composition of protein is to perform the variable length of protein sequences to fixed length feature vectors because SVM machine learning technique requires fixed length patterns. The amino acid composition is the fraction of each amino acid in a protein sequence and provides vector of 20 dimensions. The dipeptide composition was used to encapsulate the global information about each protein sequence, which gives a fixed length pattern of 400 (20×20) vectors. Both amino acids and dipeptide composition was calculated, and used as input to discrimination between cytosolic and mitochondrial tRNA synthetase using machine learning of SVM.

Composition of N-termini, C-termini and intermediate residues

We calculated separate amino acid compositions for each approach. We have used first 40 residues for N-terminus, last 60 residues for C-termini and remaining residues for the intermediate approach. For each approach separate SVM-based classifier was developed using 20 dimensions of vector of amino acid composition, one for each amino acid.

Split amino acid composition

In the case of SAAC approach, we have divided each protein sequence into three parts: (a) 40 residues of the N-terminus, (b) 60 residues of the C-terminus, and (c) the intermediate region. The variable length protein sequences were represented by a fixed length pattern of 60 dimensions of vector instead of 20 in case of standard amino acid composition. The advantage of SAAC over standard amino acid composition is that it provides greater weight of compositional biasness to proteins that have a signal at either the N or C terminus.

Attribute selection method

We have selected most significant amino acids from mitochondrial and cytosolic tRNA synthetases using WEKA 3.6.0 version (Hall et al. 2009). WEKA is a package of java programs for machine learning. In this study, we used attribute evaluator for SVMAttributeEval (parameter -X 1 -Y 0 -Z 0 -P 1.0E-25 -T 1.0E-10 -C 1.0 -N 0) method with ranker (parameter -T -1.7976931348623157E308 -N -1). We have used split composition of these selected amino acids from N-termini, C-termini and intermediate regions. These split compositions were used in SA-SAAC approach.

Position-specific scoring matrix

The PSSM was generated using the PSI-BLAST search with a cut-off *E* value of 0.01 against the large databases such as the non-redundant (NR) database available at Swiss-Prot (Altschul et al. 1997). After three iterations, PSI-BLAST generates the PSSM with the highest score from multiple alignments of the high-scoring hits by calculating the position-specific scores for each position in the alignments. The matrix contains $20 \times N$ elements, where N is the length of the target sequence, and each element represents the frequency of occurrence of each of the 20 amino acids at a particular position in the

alignment. The final PSSM was normalized using a sigmoid function. SVM input requires fixed length for machine learning; we summed all of the rows in the PSSM corresponding to the same amino acid in the sequence, and then divided each element by the length of the sequence.

Support vector machines

In this study, a highly successful machine learning technique termed as a SVM was used. Machine learning of SVM is based on the structural risk minimization principle of statistics learning theory. SVMs are a set of related supervised learning methods used for classification and regression (Vapnik 1999). We can choose and optimize number of parameters and kernels (e.g. linear, polynomial, radial basis function and sigmoidal) or any user-defined kernel. In this study we implemented SVM^{light} Version 6.01 package (Joachims 1999) of SVM and learning was carried out using three different (linear, polynomial and radial basis function) kernels. SVM takes a set of feature vectors as input, along with their output, which is used for training of model. After training, learned model can be used for the prediction of unknown examples (Kumar and Raghava 2009). In this work, the SVM training has been carried out by the optimization of various parameters of different kernels and the value of the regularization parameter C. We optimized different kernels for all approaches and preliminary tests showed that the RBF kernel gives better results than other kernels. Therefore, in this work the RBF kernel was used for all the approaches.

Fivefold cross validation

Firstly, we have used mitochondrial AARSs as positive data set and cytosolic AARSs as negative data set for the development of the tools for discrimination between these two types of AARSs. Both protein sequences of positive and negative data sets were randomly divided into five parts. Each of these five sets consists of one-fifth of positive and one-fifth of negative sequences. For training, testing and evaluating our methods, we have used a fivefold cross-validation technique (Chou and Shen 2007). In this technique, the training and testing was carried out five times, each time using one distinct set for testing and the remaining four sets for training.

Evaluation parameters

The performance evaluation of method was done by calculating the sensitivity, specificity, accuracy and the MCC of the prediction. These calculations were routinely used in these type of prediction-based studies (Bhasin and Raghava 2005; Kumar et al. 2005). These parameters can be calculated using Eqs. 1–4,

$$\text{Sensitivity} = [\text{TP}/(\text{TP} + \text{FN})] \times 100 \quad (1)$$

$$\text{Specificity} = [\text{TN}/(\text{TN} + \text{FP})] \times 100 \quad (2)$$

$$\text{Accuracy} = [(\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})] \times 100 \quad (3)$$

$$\text{MCC} = \frac{(\text{TP})(\text{TN}) - (\text{FP})(\text{FN})}{\sqrt{[\text{TP} + \text{FP}][\text{TP} + \text{FN}][\text{TN} + \text{FP}][\text{TN} + \text{FN}]}} \quad (4)$$

where, TP is correctly predicted positive (mitochondrial AARSs) proteins; TN is correctly predicted negative (cytosolic AARSs) proteins; FP is wrongly predicted positive (mitochondrial AARSs) proteins; FN is wrongly predicted negative (cytosolic AARSs) proteins.

The performance of a method is an average of five sub sets, which is created by fivefold cross-validation technique. MCC is considered to the most robust parameters for the evaluation of any prediction method (Baldi et al. 2000). The MCC value of 1 corresponds to a perfect prediction, whereas 0 corresponds to a completely random prediction. All these parameters are threshold-dependent and they require proper optimization for the better performance. The complete optimization of all parameters is very important step in SVM-based machine learning. We manually optimized all parameters and selected the one, which gives best performance.

Web-server

We have developed a user friendly web-server *MARSpred* for the prediction of mitochondrial tRNA synthetases. This prediction method is freely available from http://www.imtech.res.in/raghava/marspred web-address. In this server, PHP technologies have been used to build the dynamic web interface. We also implemented our previously developed web-server *icaars* (Panwar and Raghava 2010) with *MARSpred*. It requires protein sequence in FASTA format. Firstly, *icaars* server will predict that whether protein sequence belongs to AARS or Non-AARS. If protein sequence is predicted as AARSs then *MARSpred* server will predict whether protein sequence belongs to cytosolic or mitochondrial AARSs. We have also provided our data sets and other supplementary materials, which were used for the development of *MARSpred* web-server.

Acknowledgments The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial assistance.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Antonellis A, Green ED (2008) The role of aminoacyl-tRNA synthetases in genetic diseases. Annu Rev Genomics Hum Genet 9:87–107

Baker MJ, Frazier AE, Gulbis JM, Ryan MT (2007) Mitochondrial protein-import machinery: correlating structure with function. Trends Cell Biol 17:456–464

Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424

Berg P (1961) Specificity in protein synthesis. Annu Rev Biochem 30:293–324

Bhasin M, Raghava GPS (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem 279:23262–23266

Bhasin M, Raghava GPS (2005) GPCRsclass: a web tool for classification of amine type of G-protein coupled receptors. Nucleic Acids Res 33:W143–W147

Brindefalk B, Viklund J, Larsson D, Thollesson M, Andersson SG (2007) Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. Mol Biol Evol 24:743–756

Chou KC, Shen HB (2007) Recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

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

Doolittle WF (1998) You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 14:307–311

Duchêne AM, Pujol C, Maréchal-Drouard L (2009) Import of tRNAs and aminoacyl-tRNA synthetases into mitochondria. Curr Genet 55:1–18

Español Y, Thut D, Schneider A, de Pouplana LR (2009) A mechanism for functional segregation of mitochondrial and cytosolic genetic codes. Proc Natl Acad Sci USA 106(46): 19420–19425

Garg A, Bhasin M, Raghava GPS (2005) SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search. J Biol Chem 280(15):14427–14432

Guda C, Guda P, Fahy E, Subramaniam S (2004) MITOPRED: a web server for the prediction of mitochondrial proteins. Nucleic Acids Res 32:W372–W374

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA Data Mining Software: an update. SIGKDD Explorations 11(1):10–18

Joachims T (1999) Making large-scale SVM learning particles. In: Scholkopf B, Berges C, Smola A (eds) Advances in kernel methods support vector learning. MIT Press, Cambridge, pp 42–56

Kaur H, Raghava GPS (2004) A neural network method for prediction of beta-turn types in proteins using evolutionary information. Bioinformatics 20:2751–2758

Kumar M, Raghava GPS (2009) Prediction of nuclear proteins using SVM and HMM models. BMC Bioinformatics 10:22

Kumar M, Bhasin M, Natt NK, Raghava GPS (2005) BhairPred: a webserver for prediction of beta-hairpins in proteins from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33:W154–W159

Kumar M, Verma R, Raghava GPS (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem 281(9):5357–5363

Kumar M, Gromiha MM, Raghava GPS (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8:463

Lee JW, Beebe K, Nangle LA, Jang J, Longo-Guess CM, Cook SA, Muriel TD, Sundberg JP, Schimmel P, Ackerman SL (2006) Editing-defective tRNA synthetase causes protein misfolding and neurodegeration. Nature 443:50–55

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large datasets of proteins or nucleotide sequences. Bioinformatics 22:1658–1659

Miyaki M, Iijima T, Shiba K, Aki T, Kita Y, Yasuno M, Mori T, Kuroki T, Iwama T (2001) Alterations of repeated sequences in 5′ upstream and coding regions in colorectal tumors from patients with hereditary nonpolyposis colorectal cancer and Turcot syndrome. Oncogene 20:5215–5218

Panwar B, Raghava GPS (2010) Prediction and classification of aminoacyl tRNA synthetases using PROSITE domains. BMC Genomics 11:507

Park SG, Schimmel P, Kim S (2008) Aminoacyl tRNA synthetases and their connections to disease. Proc Natl Acad Sci USA 105:11043–11049

Raghava GPS, Han JH (2005) Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. BMC Bioinformatics 6:59

Rajbhandary UL (1997) Once there were twenty. Proc Natl Acad Sci USA 94:11761–11763

Scheper GC, van der Klok T, van Andel RJ, van Berkel CG, Sissler M, Smet J, Muravina TI, Serkov SV, Uziel G, Bugiani M, Schiffmann R, Krägeloh-Mann I, Smeitink JA, Florentz C, Van Coster R, Pronk JC, van der Knaap MS (2007) Mitochondrial aspartyl-tRNA synthetase deficiency causes leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation. Nat Genet 39:534–539

Schimmel P (2008) Development of tRNA synthetases and connection to genetic code and disease. Protein Sci 17:1643–1652

Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4(6):1581–1590

t Hart LM, Hansen T, Rietveld I, Dekker JM, Nijpels G, Janssen GM, Arp PA, Uitterlinden AG, Jørgensen T, Borch-Johnsen K, Pols HA, Pedersen O, van Duijn CM, Heine RJ, Maassen JA (2005) Evidence that the mitochondrial leucyl tRNA synthetase (LARS2) gene represents a novel type 2 diabetes susceptibility gene. Diabetes 54:1892–1895

Unseld M, Marienfeld JR, Brandt P, Brennicke A (1997) The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366, 924 nucleotides. Nat Genet 15:57–61

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999