Multi-relational Data Mining for Tetratricopeptide Repeats (TPR)-Like Superfamily Members in *Leishmania spp.*: Acting-by-Connecting Proteins

Karen T. Girão, Fátima C.E. Oliveira, Kaio M. Farias, Italo M.C. Maia,
Samara C. Silva, Carla R.F. Gadelha, Laura D.G. Carneiro, Ana C.L. Pacheco,
Michel T. Kamimura, Michely C. Diniz, Maria C. Silva, and Diana M. Oliveira

Núcleo Tarcisio Pimenta de Pesquisa Genômica e Bioinformática – NUGEN, Faculdade de
Veterinária, Universidade Estadual do Ceara – UECE, Av. Paranjana, 1700 – Campus do
Itaperi, Fortaleza, CE 60740-000 Brazil
diana.magalhaes@uece.br

Abstract. The multi-relational data mining (MRDM) approach looks for patterns that involve multiple tables from a relational database made of complex/structured objects whose normalized representation does require multiple tables. We have applied MRDM methods (relational association rule discovery and probabilistic relational models) with hidden Markov models (HMMs) and Viterbi algorithm (VA) to mine tetratricopeptide repeat (TPR), pentatricopeptide (PPR) and half-a-TPR (HAT) in genomes of pathogenic protozoa *Leishmania*. TPR is a protein-protein interaction module and TPR-containing proteins (TPRPs) act as scaffolds for the assembly of different multiprotein complexes. Our aim is to build a great panel of the TPR-like superfamily of *Leishmania*. Distributed relational state representations for complex stochastic processes were applied to identification, clustering and classification of *Leishmania* genes and we were able to detect putative 104 TPRPs, 36 PPRPs and 08 HATPs, comprising the TPR-like superfamily. We have also compared currently available resources (Pfam, SMART, SUPER-FAMILY and TPRpred) with our approach (MRDM/HMM/VA).

Keywords: Multi-relational data mining; hidden Markov models, Viterbi algorithm, tetratricopeptide repeat motif, *Leishmania* proteins.

1 Introduction

Early efforts in bioinformatics concentrated on finding the internal structure of individual genome-wide data sets; with the explosion of the 'omics' technologies, comprehensive coverage of the multiple aspects of cellular/organellar physiology is progressing rapidly, generating vast amounts of data on mRNA profiles, protein/ metabolic abundances, and protein interactions encompassing a systems-level approach that requires integrating all of the known properties of a given class of components (e.g., protein abundance, localization, physical interactions, etc.) with computational methods able to combine large and heterogeneous sets of data [1]. A

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 359–372, 2008.

technique for generation of unified mechanistic models of cellular/organellar processes (a major challenge for all who seek to discover functions of many yet unknown genes) is the multi-relational data mining (MRDM) approach, which looks for patterns that involve multiple input tables (relations) from a relational database (db) made of complex/structured objects whose normalized representation requires multiple tables [2]. MRDM extends association rule mining to search for interesting patterns among data in multiple tables rather than in one input table [3]. We have applied MRDM methods (relational association rule discovery – RARD and probabilistic relational models - PRMs) combined with hidden Markov models (HMMs) [4-5] and the Viterbi algorithm (VA) [6] to mine the tetratricopeptide repeat (TPR) [7-9] and related motifs (pentatricopeptide repeat (PPR) [10-11] and half-a-TPR (HAT) [12] in pathogenic protozoa *Leishmania spp.*. Our aim is to build a great panel of the TPR-like superfamily of proteins, whose members can be further assigned functional roles in terms of containing motifs. TPR motifs were originally identified in yeast as protein-protein interaction (PPI) modules [7], but now they are known to occur in a wide variety of proteins (over 12,000 as included in SMART nrdb) present in prokaryotic and eukaryotic organisms [8], being involved in protein-protein and protein-lipid interactions in cell cycle regulation, chaperone function and post-translation modifications [7-9]. TPRs exhibit a large degree of sequence diversity and structural conservation (two antiparallel alpha-helices separated by a turn) that might act as scaffolds for the assembly of different multiprotein complexes [13] including the peroxisomal import receptor and the NADPH oxidase [14]. Similar to TPR, PPR and HAT motifs also have repetitive patterns characterized by tandem array of repeats, where the number of motifs seems to influence the affinity and specificity of the repeat-containing protein for RNA [12,15-16]. PPR-containing proteins (PPRPs) occur predominantly in eukaryotes [10] (particularly abundant in plants), while it has been suggested that each of the highly variable PPRPs is a gene-specific regulator of plant organellar RNA metabolism. HAT repeats are less abundant and HAT-containing proteins (HATPs) appear to be components of macromolecular complexes that are required for RNA processing [10-12,15-16].

TPR-containing proteins (TPRPs) have recently attracted interest because of their versatility as scaffolds for the engineering of PPIs [17-18] and, since they are characterized by homologous, repeating structural units, which stack together to form an open-ended superhelical structure, such an arrangement is in contrast to the structure of most proteins, which fold into a compact shape [19]. The curvature created by the superhelical nature predetermines the target proteins that can bind to them [20]. TPRs, PPRs and HAT (all together referred as TPR-like motifs), form a large superfamily or the clan TPR-like [7-16]. Homologous structural repeat units are often highly divergent at the sequence level, a feature that makes their prediction challenging. Currently, several web-based resources are available for the detection of TPRs, including Pfam [21], SMART [22], and SUPERFAMILY [23], which use HMM profiles constructed from the repeats trusted to belong to the family (from closely homologous repeats); therefore, divergent repeat units often get a negative score and are not considered in computing the overall statistical significance, even though they are individually significant [18]. For this reason Pfam, SMART, and SUPERFAMILY perform with limited accuracy in detecting remote homologs of known TPRPs and in delineating the individual repeats within a protein [18]. A new profile-based method [18], TPRpred, uses a P-value- dependent score offset to include divergent repeat units and to exploit the tendency of repeats to occur in tandem. Although TPRpred indeed performs significantly better in detecting divergent repeats in TPRPs, and finds more individual repeats than the afore mentioned methods, we have noticed that it still fails to detect some particular groups of members of TPR-like superfamily, such as now we demonstrate for *Leishmania spp.* Since the characterization of proteins of a given family often relies on the detection of regions of their sequences shared by all family members, while computing the consensus of such regions provides a motif that is used to recognize new members of the family, our approach of HMMs/VA with MRDM was suitable to detect 104 TPRPs, 36 PPRPs and 08 HATPs in *Leishmania spp.* genomes, a greater number than Pfam, SMART, SUPERFAMILY are able to yield (Table 1) and slightly higher than TPRpred.

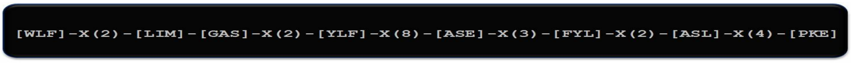
2 Methods

2.1 Data Sources and Bioinformatics Tools

We have used publicly available datasets of individual or clusters of gene/protein data on *Leishmania spp.*, mainly *L. major, L. braziliensis*, *L. infantum* and related trypanosomatids (GeneDB [24] and NCBI/Entrez - www.ncbi.nlm.nih.gov/sites/gquery). Variants of BLAST [25] and GlimmerHMM [26] were widely used for sequence similarity searches, comparisons and gene predictions. External db searches were performed against numerous collections of protein motifs and families. Gene ontology (GO) terms were assigned, based on top matches to proteins with GO annotations from Swiss-Prot/trEMBL (www.expasy.org/sprot) and AMIGO after GeneDB (www.genedb.org/amigo/perl) access. Functional assignment of genes/gene products was inferred using the RPS-BLAST search against conserved domain db (CDD) [27]. For protein domain identification and analysis of protein domain architectures, Simple Modular Architecture Research Tool (SMART) [22], Pfam [21], SUPERFAMILY [23] and TPRpred [18] were used. For multiple alignments we used MUSCLE [28].

2.2 Finding a TPR-Like Regular Expression

TPR motif sequence is loosely based around the consensus residues -W-LG-Y-A-F-A-P-. TPRs are minimally conserved (degenerate and variable) regions of 34-residue long extension (with exceptions accepted to the range of 31 residues [14]). Three-dimensional structural data have shown that tandem arrays of 3-16 TPR motifs generate a right-handed helical structure with an amphipathic channel that might accommodate the complementary region of a target protein [7, 9, 14]. The PPR motif is a degenerate 35-residue sequence, closely related to the 34-residue TPR motif. On the basis of the solved structure of a TPR domain [9] as well as modeling approaches [10], each PPR domain is though to be configured also as two distinct antiparallel alpha-helices, helices A and B. In PRPPs, 2-26 tandem repeats of these alpha-helical pairs are predicted to form a superhelix that encloses a central spiral groove with a positively charged ligand-binding surface [10]. Although there exists no position characterized by an invariant residue, a consensus sequence pattern of small and large hydrophobic residues has been defined: small hydrophobic residues are commonly observed at positions 8, 20, and 27, while large ones are at 4, 17, and 24 [14]. The consensus sequence for TPR-like motif is given below (1) and it has been used as a regular expression, which defines the most probable amino acid (aa) at each position within this core, to fully exploit the TPR motif finding in *Leishmania spp*. genomes. As reported in [29], we systematically solved inconsistencies in the motif annotation by manual expertise. Since motif occurrences are adjacent in sequences, we could define the motif sequence of a protein as the succession of motifs read from the N toward the C terminus.



```
[WLF]-X(2)-[LIM]-[GAS]-X(2)-[YLF]-X(8)-[ASE]-X(3)-[FYL]-X(2)-[ASL]-X(4)-[PKE]
```

(1)

2.3 Definition of a TPR-Like Protein

We have defined a TPR-like protein as any protein sequence containing a TPR-like motif that fits in our regular expression (1), which also, by reference, confirms to a set of known bona fide domains contained in TPR-like superfamily [a.118.8] of SCOP (v.1.69) [18,30], SMART (v.5.0) [22], TPRpred [18] and SUPERFAMILY [23]. Classification criteria are supported by structural/sequence similarity, plus searches with remote homology prediction.

2.4 Profile Generation After Querying TPR-Like Motifs

Aware that performance dependence on any sequence profiles relies on either the selectivity or sensitivity of its regime, respectively depending on the number of close or remote homologs used [18], we have established a fixed threshold value to include a minimum number of remote homologs (to avoid having too many false positives). Initial profiles were generated by iterative searches against non-redundant dbs (nrdbs) at NCBI and GeneDB, filtered to a maximum pairwise sequence identity of 60% (nr-60) by CD-HIT [31-32], slightly modified after [18] in a sense that we have extracted sequences conservatively with PSI-BLAST through multiple iterations using the TPR-like regular expression (1) as a query sequence. We, then, performed iterative searches to convergence on nr-60 minus TPRPs (detected by Pfam, SMART, SUPERFAMILY and TPRPred) with various threshold parameters to test the resulting profiles on a positive (TPR-like) or negative set (non TPR-like). Best profiles were selected based on its performance on a predicted family assignment, as illustrated on Figure 1a.

2.5 TPR-Like Superfamily Assignment

To provide structural (and hence implied functional) assignments to TPR-like proteins at the superfamily level, structured sequences from available *Leishmania* genomes were randomly selected and parsed into unique 24,708 sequences. Each sequence was a labeled input to a multi-class motif classifier. To pick the best method to represent one or more of the three target motifs, we compared the results of motif classifiers when the sequence was presented as a (I) TPR-containing, (II) PPR-containing (III) HAT-containing, (IV) combination of any two or three motifs, and (V) not-containing target motifs. Performance was measured by classification precision, recall and F1 measure (a composite measure of classification precision and recall).

2.6 Hidden Markov Models (HMMs) and the Viterbi Algorithm (VA)

A HMM is a probabilistic network of nodes, so called states. One state q_i is connected to another state q_j by a transition probability *ij*. Non-silent states are able to emit an alphabet of symbols [4-5]. A special topology of HMMs, termed pHMM, is frequently used in homology detection of protein families [33]. Transition and emission probabilities are estimated by a maximum likelihood approach combined with a standard dynamic programming algorithm for decoding HMMs, the Viterbi (VA) [6]) to get site and path dependent probabilities for every hidden state in the posterior decoding. In a first validation step we used the feature of trained HMMs to emit domain-specific sequences according to their model parameters. Sequences were compared with generated state paths in the same way as described earlier [29]. The process of generation was repeated 10 times for every TPR-like motif. To fully exploit the sequential ordering of motifs in a set, we used pHMMs to label motif types. We have transformed the motif categorization problem into a HMM sequence alignment problem. The HMM states correspond to the motif types. Labeling motifs in a sequence is equivalent to aligning the sequences to HMM states. There are five states in our HMM model: (I) TPR-containing, (II) PPR-containing (II) HAT-containing, (IV) combination of any two or all motifs, and (V) not-containing target motifs. Transition probabilities between these states were estimated from the training data by dividing the number of times each transition occurs in the training set by the sum of all the transitions. The state emission probabilities were calculated from the score output reported by the multi-class classifiers. Given the HMM model [33], state emission probabilities and state transition probabilities, VA was used to compute most-likely sequence of states that emit (any of the target) motifs in sequences. Subsequently, the state associated with the motif was extracted from the most-likely sequence of states [34].

2.7 Multi-relational Data Mining (MRDM) Method

Algorithms for RARD are well suited for exploratory data mining due to the flexibility required to experiment with examples more complex than feature vectors and patterns more complex than item sets [35], such as the case with TPR-like motifs. An adequate approach of machine learning [36] focuses on learning a complex web of relationships among a collection of diverse objects rather than supervised learning from independent and identically distributed training examples (a classifier f that given an object x would produce as output a classification label $y = f(x)$). Such formalism, developed as PRMs [36-37], can represent these webs of relationships and support learning and reasoning with them [38]. PRMs are a multi-relational form of Bayesian networks that allow descriptions of a template for a probability distribution. This, together with a set of motif objects, defines a distribution over the attributes of the objects. Such a model can then be used for reasoning about an entity using the entire rich structure of knowledge encoded by the relational representation [37,39]. For each PRM, we were interested in constructing a model whose trades off fit to data with the TPR-like motif model complexity. This tradeoff allows us to avoid fitting the training data too closely, which would reduce our ability to predict unseen data.

3 Results and Discussion

3.1 TPR-Like Motif Localization Task

As illustrated on Fig. 1, we have applied two MRDM methods (RARD and PRMs) after HMMs/VA to mine TPR, PPR and HAT repeats in protein sequences of *Leishmania spp*. Provided six variants of the data set for the TPR-like motif localization task (considering four TPRs, one PPR and one HAT in the TPR-like clan), the first version consisted of a single table with 24,708 attributes and the second consisted of two tables with 26 attributes in total. We used a normalized version of the data set with two tables. The names of the two original tables are *motifs_relation* and *interactions_relation*. The *motifs_relation* table contained 120 different motifs but there could be more than one row in the table for each motif. The attribute *motif_id* identifies a motif is uniquely. Since our current implementation of MRDM requires that the target table must have a primary key, it was necessary to normalize the *motifs_relation* table before we could use it as the target table. This normalization was achieved by creating the tables named *motif*, *interaction*, and *composition* as follows: Attributes in the *motifs_relation* table that did not have unique values for each motif were placed in *composition table* and the rest of attributes were placed in *motif* table. The *motif_id* attribute is a primary key in the *motif* table and as a foreign key in *composition* table. The *interaction* table is identical to the original *interactions_relation* table. This represents one of several ways of normalizing the original table and renormalization of the relational db has an impact on the entity-relation diagram for the renormalized version of *Motif Localization* db. Thus, for

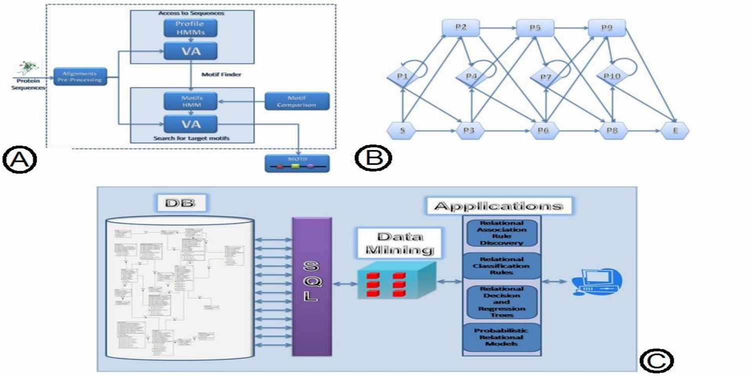


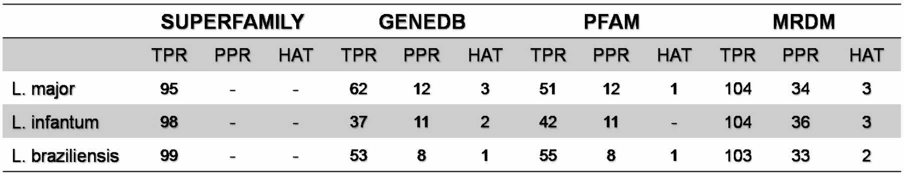
Fig. 1. Overall schema for multi-relational data mining (MRDM) approach (C) comprising the search for TPR, PPR and HAT motifs in available *Leishmania* genomes after (A) hidden Markov models (HMMs) powered by (B) the Viterbi algorithm (VA), a combined method for superfamily assignment on searches among *Leishmania* fully sequenced species and trypanosomes [24]. The input to VA is a HMM in sequences of length *L*. The output is the highest probability path through the HMM that could generate the input sequence.

TPR-like motif localization task, the target table is *motif* and the target attribute is *localization*. From this point of view, the training set consists of 120 motifs and the test set 68. The experiments described here focused on building a classifier for predicting the localization of motif-containing proteins by assigning the corresponding instance to one of six possible localizations. For this motif localization task, we have chosen to construct a classifier using all training data and test the resulting classifier on the test set provided by *Leishmania* sequences. This task presents significant challenges because many attribute values in training instances corresponding to the 120 training motifs are missing. Initial experiments using a special value to encode a missing value for an attribute resulted in classifiers whose accuracy is around 40% on the test data. This prompted us to investigate incorporation of other approaches to handling missing values. Replacing missing values by the most common value of the attribute for the class during training resulted in an accuracy of around 68%. This shows that providing reasonable guesses for missing values can significantly enhance the performance of MRDM on our data sets. However, in practice, since class labels for test data are unknown, it is not possible to replace a missing attribute value by the most frequent value for the class during testing. Hence, there is a need for better ways of handling missing values (e.g., predicting missing values based on values of which attributes?).

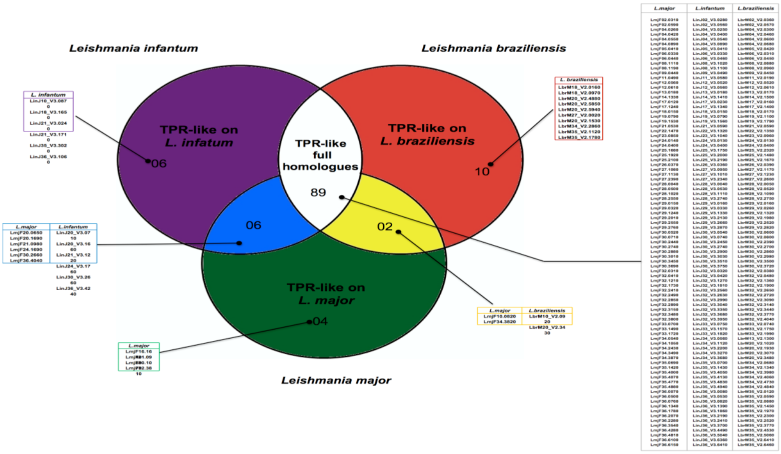
3.2 Identification of the TPR-Like Superfamily in *Leishmania* spp. Genomes

The percentage of repeat-containing proteins, such as TPR-like, grows with the complexity of the organism, with repeat proteins being particularly abundant in multicellular organisms [40]. Genomes of unicellular eukaryotes, as *Leishmania*, usually possess a relatively high number of putative encoding genes (around 8,000 genes in *L. major*, e.g.) [24]. Analyses of such a large number of coded proteins require that the characterization of a given family of proteins be dependent on detection of regions of their sequences shared by all family members. Computing the consensus of such regions provides a motif that is used to recognize new members of the family [41]. With the sequencing completion of 03 *Leishmania* and several trypanosomes genomes [24], we were able to search for all TPR-like genes in *Leishmania* using the defining characteristic of a TPR-like protein. As depicted by Tab. 1, numbers of members detected through different tools (GeneDB, Pfam and

Table 1. Comparative results of TPR-like motif finding in *Leishmania* genes obtained with three standard tools (Superfamily, GeneDB and Pfam) and with our method (MRDM/HMM/VA). Numbers are shown in terms of TPR, PPR and HAT-containing proteins in three species (*L. major*, *L. infantum* and *L. braziliensis*).



	SUPERFAMILY			GENEDB			PFAM			MRDM		
	TPR	PPR	HAT	TPR	PPR	HAT	TPR	PPR	HAT	TPR	PPR	HAT
L. major	95	-	-	62	12	3	51	12	1	104	34	3
L. infantum	98	-	-	37	11	2	42	11	-	104	36	3
L. braziliensis	99	-	-	53	8	1	55	8	1	103	33	2



L.major	L.infantum	L.braziliensis
LmjF02.0310	LinJ02_V3.0280	LbrM02_V2.0360
LmjF02.0590	LinJ02_V3.0560	LbrM02_V2.0570
LmjF04.0260	LinJ04_V3.0250	LbrM04_V2.0300
LmjF04.0420	LinJ04_V3.0400	LbrM04_V2.0460
LmjF04.0550	LinJ04_V3.0540	LbrM04_V2.0600
LmjF04.0890	LinJ04_V3.0890	LbrM04_V2.0680
LmjF05.0410	LinJ05_V3.0410	LbrM05_V2.0420
LmjF06.0330	LinJ06_V3.0330	LbrM06_V2.0310
LmjF06.0440	LinJ06_V3.0460	LbrM06_V2.0450
LmjF08.1110	LinJ08_V3.1020	LbrM08_V2.0880
LmjF08.1190	LinJ08_V3.1100	LbrM08_V2.0960
LmjF09.0440	LinJ09_V3.0490	LbrM09_V2.0450
LmjF11.0490	LinJ11_V3.0580	LbrM11_V2.0190
LmjF12.0560	LinJ12_V3.0520	LbrM12_V2.0520
LmjF12.0610	LinJ12_V3.0560	LbrM12_V2.0610
LmjF13.0180	LinJ13_V3.0180	LbrM13_V2.0170
LmjF14.1330	LinJ14_V3.1410	LbrM14_V2.1500
LmjF17.0120	LinJ17_V3.0230	LbrM17_V2.0160
LmjF17.1240	LinJ17_V3.1340	LbrM17_V2.1400
LmjF18.0150	LinJ18_V3.0150	LbrM18_V2.0170
LmjF19.0790	LinJ19_V3.0790	LbrM19_V2.1100
LmjF19.1530	LinJ19_V3.1560	LbrM19_V2.1790
LmjF21.0530	LinJ21_V3.0590	LbrM21_V2.0590
LmjF22.1470	LinJ22_V3.1320	LbrM22_V2.1350
LmjF23.0850	LinJ23_V3.1040	LbrM23_V2.0960
LmjF24.0140	LinJ24_V3.0130	LbrM24_V2.0130
LmjF24.0400	LinJ24_V3.0400	LbrM24_V2.0400
LmjF25.1680	LinJ25_V3.1750	LbrM25_V2.2320
LmjF25.1920	LinJ25_V3.2000	LbrM25_V2.1480
LmjF25.2100	LinJ25_V3.2190	LbrM25_V2.1670
LmjF26.0370	LinJ26_V3.0360	LbrM26_V2.0390
LmjF27.1080	LinJ27_V3.0950	LbrM27_V2.1170
LmjF27.1130	LinJ27_V3.1010	LbrM27_V2.1230
LmjF27.2390	LinJ27_V3.2340	LbrM27_V2.2600
LmjF28.0040	LinJ28_V3.0040	LbrM28_V2.0050
LmjF28.0500	LinJ28_V3.0530	LbrM28_V2.0520
LmjF28.1020	LinJ28_V3.1110	LbrM28_V2.1090
LmjF28.2550	LinJ28_V3.2740	LbrM28_V2.2750
LmjF29.0150	LinJ29_V3.0160	LbrM29_V2.0160
LmjF29.0320	LinJ29_V3.0330	LbrM29_V2.0260
LmjF29.1240	LinJ29_V3.1330	LbrM29_V2.1320
LmjF29.2010	LinJ29_V3.2130	LbrM29_V2.1980
LmjF29.2550	LinJ29_V3.2660	LbrM29_V2.2520
LmjF29.2760	LinJ29_V3.2870	LbrM29_V2.2820
LmjF30.0520	LinJ30_V3.0540	LbrM30_V2.0600
LmjF30.0710	LinJ30_V3.0740	LbrM30_V2.0800
LmjF30.2440	LinJ30_V3.2450	LbrM30_V2.2390
LmjF30.2740	LinJ30_V3.2740	LbrM30_V2.2700
LmjF30.2880	LinJ30_V3.2900	LbrM30_V2.2860
LmjF30.3010	LinJ30_V3.3030	LbrM30_V2.2980
LmjF30.3450	LinJ30_V3.3510	LbrM30_V2.3500
LmjF30.3690	LinJ30_V3.3750	LbrM30_V2.3720
LmjF32.0310	LinJ32_V3.0320	LbrM32_V2.0380
LmjF32.0410	LinJ32_V3.0420	LbrM32_V2.0480
LmjF32.1210	LinJ32_V3.1270	LbrM32_V2.1360
LmjF32.1730	LinJ32_V3.1810	LbrM32_V2.1900
LmjF32.2410	LinJ32_V3.2560	LbrM32_V2.2650
LmjF32.2490	LinJ32_V3.2630	LbrM32_V2.2720
LmjF32.2850	LinJ32_V3.2990	LbrM32_V2.3090
LmjF32.2890	LinJ32_V3.3040	LbrM32_V2.3140
LmjF32.3150	LinJ32_V3.3350	LbrM32_V2.3440
LmjF32.3480	LinJ32_V3.3680	LbrM32_V2.3770
LmjF32.3800	LinJ32_V3.3950	LbrM32_V2.4040
LmjF33.0700	LinJ33_V3.0750	LbrM33_V2.0740
LmjF33.1490	LinJ33_V3.1570	LbrM33_V2.1750
LmjF33.1720	LinJ33_V3.1820	LbrM33_V2.1990
LmjF34.0540	LinJ34_V3.0560	LbrM13_V2.1300
LmjF34.1060	LinJ34_V3.1120	LbrM20_V2.1020
LmjF34.2430	LinJ34_V3.2200	LbrM20_V2.1930
LmjF34.3490	LinJ34_V3.3270	LbrM20_V2.3070
LmjF34.3870	LinJ34_V3.3680	LbrM20_V2.3480
LmjF35.0690	LinJ35_V3.0700	LbrM34_V2.0680
LmjF35.1420	LinJ35_V3.1430	LbrM34_V2.1340
LmjF35.4000	LinJ35_V3.4050	LbrM34_V2.3980
LmjF35.4070	LinJ35_V3.4130	LbrM34_V2.4060
LmjF35.4770	LinJ35_V3.4830	LbrM34_V2.4730
LmjF35.4880	LinJ35_V3.4940	LbrM34_V2.4840
LmjF36.0070	LinJ36_V3.0080	LbrM35_V2.0120
LmjF36.0500	LinJ36_V3.0530	LbrM35_V2.0590
LmjF36.0760	LinJ36_V3.0820	LbrM35_V2.0880
LmjF36.1340	LinJ36_V3.1390	LbrM35_V2.1450
LmjF36.1780	LinJ36_V3.1860	LbrM35_V2.1970
LmjF36.2070	LinJ36_V3.2190	LbrM35_V2.2300
LmjF36.2280	LinJ36_V3.2410	LbrM35_V2.2520
LmjF36.3540	LinJ36_V3.3700	LbrM35_V2.3770
LmjF36.4280	LinJ36_V3.4490	LbrM35_V2.4530
LmjF36.4810	LinJ36_V3.5040	LbrM35_V2.5060
LmjF36.6100	LinJ36_V3.6360	LbrM35_V2.6410
LmjF36.6150	LinJ36_V3.6410	LbrM35_V2.6460

Fig. 2. Schematic diagram of TPR-containing genes in genomes of *Leishmania major, L. infantum* and *L. braziliensis*, detected after a multi-relational data mining and hidden Markov model/Viterbi algorithm approach (MRDM/HMM/VA). Shared ortologues among the three species are illustrated as colored circles intersections and individual identifiers (GeneDB IDs) for putative TPRPs are given as lateral tables.

Superfamily) are shown in comparison to our method (MRDM/HMM/VA), which is able to assign a significantly larger number of TPR-like motif-containing proteins in *Leishmania*: 104 TPRs and 36 PPRs at the most and 08 HATs in total. These members are elements putatively involved in several key cellular processes, such as glycosome biogenesis (PEX5 and PEX14) and flagellar pathways (IFT subunits, cyclophilins, phosphatases), besides binding partners of either motor or cargo proteins (kidins220/ARMS and other members of the KAP family of P-loop NTPases) or those involved with assembly/disassembly of protein complexes. The resulting descriptions of the families and its members, a good example of relevant patterns found along with reasonable assignment of family members with our approach, should provide a solid and unified platform on which future genetic and functional studies regarding *Leishmania* TPRPs can be based.

3.3 TPR-Encoding Genes in *Leishmania spp*. Genomes

We first used the alignment of 275 sequences previously identified as putative strict TPR-containing motifs (obtained from Superfamily, SMART, Pfam and TPRpred) to obtain the consensus model (Fig. 3). This TPR signature matrix was subsequently used to search for TPR motifs in the six reading frames of whole *Leishmania* genomes. Multiple alignment of *Leishmania* TPRP sequences revealed that most of substitutions in the TPRs occur at nonconsensus positions; consensus residues are selectively conserved between ortologues (particularly in *Trypanosoma spp*). Because TPR motifs are highly degenerate, a fairly large number of false positive hits were expected. However, because TPR motifs appear usually as tandem repeats, we could remove most random uninteresting matches by omitting all orphan TPR motifs that

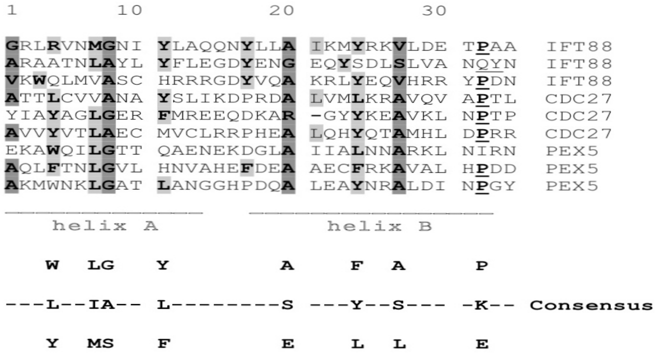

1 10 20 30
GRLRVNMGNI YLAQQNYLLA IKMYRKVLDE TPAA IFT88
ARAATNLAYL YFLEGDYENG EQYSDLSLVA NQYN IFT88
VKWQLMVASC HRRRGDYVQA KRLYEQVHRR YPDN IFT88
ATTLCVVANA YSLIKDPRDA LVMLKRAVQV APTL CDC27
YIAYAGLGER FMREEQDKAR -GYYKEAVKL NPTP CDC27
AVVYVTLAEC MVCLRRPHEA LQHYQTAMHL DPRR CDC27
EKAWQILGTT QAENEKDGLA IIALNNARKL NIRN PEX5
AQLFTNLGVL HNVAHEFDEA AECFRKAVAL HPDD PEX5
AKMWNKLGAT LANGGHPDQA LEAYNRALDI NPGY PEX5
helix A
helix B
W LG Y A F A P
---L--IA-- L--------S ---Y--S--- -K-- Consensus
Y MS F E L L E

Fig. 3. Multiple sequence alignment of typical TPR motifs present in IFT88, CDC27 and PEX5 proteins of *Leishmania* (LmjF27.1130, LmjF05.0410 and LmjF35.1420). TPR motif residues are shown with *dark* and *light gray shaded boxes* for small and large hydrophobic residues, respectively. Small hydrophobic residues are commonly observed at positions 1, 8, 20 and 27. Position 32 is frequently proline (*bold underlined*), located at the C terminus of helix B, and large hydrophobic residues are also located at particular positions, especially 4, 17, and 24. Schematic consensus for TPR is illustrated.

were found farther than 200 nucleotides from any other TPR motif. The 465 TPR motifs retained formed 132 clusters, each of which comprised a putative TPR gene. Each TPR motif cluster was then investigated in detail by manually analyzing the positions and reading frames of the TPR motifs compared with available *Leishmania* genomes (1) open reading frame (ORF) models and (2) predicted protein sequences within potential coding sequences. From this analysis, 104 putative TPR ORF models were constructed (i.e., 28 motif clusters were discarded or fused with other clusters). TPR genes are fairly evenly distributed throughout the 36 mini-chromosomes of *L. major*, with little in the way of obvious clusters. The densest grouping of TPR genes lies on chromosomes 30, 32 and 36, the latter which contains 13 genes, the maximum number found in any isolate chromosome of *L. major*.

3.4 Functional Features of TPR-Like Motifs

The preliminary functional predictions of a range of family members performed here, together with the sparse data on these proteins in *Leishmania* published so far, allows us to propose putative models in which TPR-like proteins might play the role of sequence-specific adaptors for a variety of other RNA-associated proteins. Such models, yet requiring further testable hypothesis, can surround a testable prediction: that TPR-like proteins in *Leishmania* might be directly or indirectly associated with specific RNA sequences and with defined effector proteins, as previously suggested in *Arabidopsis* [15]. Future work needs to be directed toward the identification of these factors to elucidate the precise functions of one of the largest and least understood protein families in *Leishmania*, the TPR-like. For now, our MRDM approach may be also relevant for other families of proteins with repeated motifs, in a similar way to what was reported by [42]. We must recall that the *L. major* genome contains 708 predicted proteins annotated with the term *repeat* in their descriptions [24], including only 62 out of the 104 TPRPs and 12 out of the 36 PPRPs that we have identified here. For instance, there are 18 proteins containing repeats of the Kelch motif often associated to a F-box domain, 169 WD40 repeat-containing proteins and 121 proteins with Leu-rich repeats frequently associated to a protein kinase domain. Others cases are armadillo (61) and ankyrin (45) repeats-containing proteins. In some of these protein families, the region containing the repeats is a large part of the proteins that can, and should be, a valuable target for applying MRDM methods.

3.5 PPR- and HAT-Encoding Genes in *Leishmania spp*. Genomes

The name PPR was coined based on its similarity to the better-known TPR motif [10]. PPRPs make up a significant proportion of the unknown function proteins in many organisms, but only few of them have functional roles ascribed, although a putative RNA-binding function is widely accepted [15] and one PPRP is involved in RNA editing [11]. The existence of a large family of PRPPs only became apparent with the *Arabidopsis* Genome Initiative that revealed 446 PPR coding genes – 6% of its entire genome [15]. The PPRP family has been divided, on the basis of their motif content and organization, into two subfamilies: the PPRP-P and the exclusive plant combinatorial and modular proteins (PCMPs). PPR motifs have been found in all eukaryotes analyzed to date, but with an extraordinary discrepancy in numbers between plant and nonplant organisms (the human genome encodes only six putative PPRPs). Trypanosomatids and other flagellated organisms are expected to have an intermediate number (around one hundred PPR genes), a number still far from the 36 PPR-encoding genes we have found here (12 of them annotated at GeneDB as conserved hypothetical proteins of *L. major*). Recent reports [43-44] mention more than twenty (respectively 23 and 28) PPRPs identified in *Trypanosoma brucei* (with at least 25 ortologues found in *L. major*) and with a predicted indication that most of these proteins are targeted to mitochondria. As of Release 2.1 of GeneDB [24] with curated annotations of *Leishmania* genes, 13 of the GeneDB ORFs are annotated as conserved hypothetical proteins that contain PPR motifs based on matches with the PFAM profile PF01535 or SMART profile IPR002885. None of *Leishmania* GeneDB models are annotated as homologs of known PPRPs. Of the two sets of ORF models (ours and GeneDB's), 12 are identical (i.e., our analysis agreed with the GeneDB model). The 13th GeneDB model does not have an equivalent in our set because we did not consider it to be a PPRP by our criteria (lacking tandem motifs matching our HMM profiles) Twenty-one of our models have no GeneDB equivalent and correspond to genes apparently overlooked during annotation or considered to be pseudogenes. In all, 22 of our 34 models differ in at least some respects from the corresponding GeneDB model, but correspond quite well to the 28 PPRPs identified in *T brucei* [44], what reinforces how well conserved PPR genes seem to be in trypanosomatids. It should be noted that in very few of these cases are molecular data available that can be used to decide between discordant models. Our choice has been generally made by comparison with other genes in the family and a general familiarity with these proteins. A noticeable characteristic of PPR genes is that they rarely contain introns within coding sequences even in higher eukaryotes (more than 80% of known PPR genes of plants unexpectedly do not contain introns), what is also true for *Leishmania* ORF models (an obvious extension for trypanosomatid genes that usually do not contain introns anyway). This characteristic might explain why PPR genes are relatively short (on average <2 kb) despite the fact that PPRPs are comparatively large proteins (680 aa on average).

HAT repeats have three aromatic residues with a conserved spacing, being structurally and sequentially similar to TPRs/PPRs, although they lack the highly conserved alanine and glycine residues found in TPRs. The number of HAT repeats found in different proteins varies between 9 to 12. HATPs appear to be components of macromolecular complexes that are required for RNA processing and the HAT motif has striking structural similarities to HEAT repeats (IPR000357), being of a similar length and consisting of two short helices connected by a loop domain, as in HEAT repeats [10-12, 15-16]. Our survey identified a total of 08 putative HATPs (Table 1) in the three species of *Leishmania* analyzed, but the lack of general information on HATPs does not allow any further indication on their definite significance on the protozoan genome. The detection of such a small, but significant, presence of HATPs in *Leishmania* is certainly an issue for future investigation.

4 Conclusions

We have performed bioinformatics analyses of *Leishmania* TPR, PPR and HAT proteins with an integrated MRDM/HMM/VA approach that, in contrast to other currently available resources (PFAM, SMART, SUPERFAMILY, TPRpred), seeks to capture as much model information as possible in the pattern matching heuristic, without resorting to more standard motif discovery methods. TPR genes are ubiquitous, whereas PPRs and HATs are mostly found in eukaryotes, but, in common, they have the fact of being largely unexplored in *Leishmania* parasites. Diffusion of new developments and applications of MRDM techniques to data-driven knowledge discovery problems in bioinformatics is a future direction towards better power of biological inference after sequence and structural analyses.

References

1. Ideker, T., Bafna, V., Lemberger, T.: Integrating scientific cultures. Mol. Syst. Biol. 3, 105–112 (2007), doi:10.1038/msb4100145
2. Getoor, L.: Multi-relational data mining using probabilistic relational models: research summary. In: Knobbe, A.J., van der Wallen, D.M.G. (eds.) Proceedings 1st Workshop in Multi-relational Data Mining, KDD (2001)
3. Dehaspe, L., De Raedt, L.: Mining association rules in multiple relations. In: Džeroski, S., Lavrač, N. (eds.) ILP 1997. LNCS, vol. 1297. Springer, Heidelberg (1997)
4. Eddy, S.R.: Profile hidden Markov models. Bioinformatics 14, 755–763 (1998)
5. Winters-Hilt, S.: Hidden Markov Model Variants and their Application. BMC Bioinformatics 7, 14 (2006)
6. Forney Jr., G.D.: The Viterbi algorithm. Proc. IEEE 61, 268 (1973)
7. Blatch, G.L., Lässle, M.: The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. Bio. Essays 21, 932–939 (1999)
8. D'Andrea, L.D., Regan, L.: TPR proteins: the versatile helix. Trends Biochem. Sci. 28, 655–662 (2003)

9. Das, A.K., Cohen, P.W., Barford, D.: The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. EMBO. J. 17, 1192–1199 (1998)
10. Small, I.D., Peeters, N.: The PPR motif – a TPR-related motif prevalent in plant organellar proteins. Trends Biochem. Sci. 25, 46–47 (2000)
11. Kotera, E., Tasaka, M., Shikanai, T.: A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. Nature 433, 326–330 (2005)
12. Preker, P.J., Keller, W.: The HAT helix, a repetitive motif implicated in RNA processing. Trends Biochem. Sci. 23, 15–16 (1998)
13. Scheufler, C., Brinker, A., Bourenkov, G., et al.: Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. Cell 101, 199–210 (2000)
14. Koga, H., Terasawa, H., Nunoi, H., et al.: Tetratricopeptide Repeat (TPR) Motifs of p67phox Participate in Interaction with the Small GTPase Rac and Activation of the Phagocyte NADPH Oxidase. Biol. Chem. 274, 25051–25060 (1999)
15. Lurin, C., Andrés, C., Aubourg, S., et al.: Genome-Wide Analysis of Arabidopsis Pentatricopeptide Repeat Proteins Reveals Their Essential Role in Organelle Biogenesis. The Plant Cell 16, 2089–2103 (2004)
16. Rivals, E., Bruyère, C., Toffano-Nioche, C., Lecharny, A.: Formation of the Arabidopsis Pentatricopeptide Repeat Family. Plant Physiol. 141, 825–839 (2006)
17. Main, E.R.G., Lowe, A.R., Mochrie, S.G.J., Jackson, S.E., Regan, L.: A recurring theme in protein engineering: the design, stability and folding of repeat proteins. Curr. Opin. Struct. Biol. 15, 464–471 (2005)
18. Karpenahalli, M.R., Lupas, A.N., Söding, J.: TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. BMC Bioinformatics 8, 2 (2007), doi:10.1186/1471-2105-8-2
19. Groves, M.R., Barford, D.: Topological characteristics of helical repeat proteins. Curr. Opin. Struct. Biol. 9, 383–389 (1999)
20. Kobe, B., Kajava, A.V.: When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem. Sci. 25, 509–515 (2000)
21. Sonnhammer, E.L., Eddy, S.R., Durbin, R.: Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405–420 (1997)
22. Schultz, J., Milpetz, F., Bork, P., Ponting, C.P.: SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. USA 95, 5857–5864 (1998)
23. Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., Gough, J.: The superfamily database in 2004: additions and improvements. Nucleic Acids Res. 32, 235–239 (2004)
24. Hertz-Fowler, C., Peacock, C.S., Wood, C., et al.: GeneDB: a resource for prokaryotic and eukaryotic organisms. Nucleic Acids Res. 32, D339–D343 (2004) The Pathogen Sequencing Unit - Wellcome Trust Sanger Institute – GeneDB – (2004), `http://www.genedb.org`
25. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
26. Majoros, W.H., Pertea, M., Salzberg, S.L.: TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878–2879 (2004)
27. Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., et al.: CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35, D237–240 (2007)

28. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004)
29. Pacheco, A.C.L., Araujo, F.F., Kamimura, M.T., et al.: Following the Viterbi Path to Deduce Flagellar Actin-Interacting Proteins of Leishmania spp.: Report on Cofilins and Twinfilins. In: Pham, T. (ed.) AIP Proceedings of Computer Models for Life Sciences, CMLS 2007, vol. 952, pp. 315–324. American Institute of Physics, Australia (2007)
30. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540 (1995)
31. Li, W., Jaroszewski, L., Godzik, A.: Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283 (2001)
32. Li, W., Jaroszewski, L., Godzik, A.: Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics 18, 77–82 (2002)
33. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846–856 (1998)
34. Friedrich, T., Pils, B., Dandekar, T., Schultz, J., Müller, T.: Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics 22, 2851–2857 (2006), doi:10.1093/bioinformatics/btl486
35. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1300–1307. Morgan Kaufman, Stockholm (1999)
36. Getoor, L., Friedman, N., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Dzeroski, S., Lavrac, N. (eds.) Relational Data Mining, pp. 307–335. Kluwer, Dordrecht (2001)
37. Getoor, L., Taskar, B., Koller, D.: Using probabilistic models for selectivity estimation. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 461–472. ACM Press, New York (2001)
38. Craven, M., Page, D., Shavlik, J., Bockhorst, J., Glasner, J.: A probabilistic learning approach to whole-genome operon prediction. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pp. 116–127. AAAI Press, La Jolla (2000)
39. Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D.: Rich probabilistic models for gene expression. Bioinformatics 1, 1–10 (2001)
40. Bjorklund, A.K., et al.: Expansion of protein domain repeats. PLoS Comput. Biol. 2, 114 (2006)
41. Servant, F., Bru, C., Carrère, S., et al.: ProDom: automated clustering of homologous domains. Brief. Bioinform. 3, 246–251 (2002)
42. Rivals, E., Bruyere, E., Toffano-Nioche, C., Lecharny, A.: Formation of the Arabidopsis pentatricopeptide repeat family. Plant Physiol. 141, 825–839 (2006)
43. Mingler, M.K., Hingst, A.M., Clement, S.L., et al.: Identification of pentatricopeptide repeat proteins in Trypanosoma brucei. Mol. Biochem. Parasitol. 150, 37–45 (2006)
44. Pusnik, M., Small, I., Read, L.K., Fabbro, T., Schneider, A.: Pentatricopeptide Repeat Proteins in Trypanosoma brucei Function in Mitochondrial Ribosomes. Mol. Cell. Biol. 27, 6876–6888 (2007)
45. NCBI (National Center for Biotechnology Information / Entrez / Cn3D (All Databases), http://www.ncbi.nlm.nih.gov/sites/gquery
46. Swiss-Prot/trEMBL, http://www.expasy.org/sprot
47. AMIGO after GeneDB access, http://www.genedb.org/amigo/perl

48. SMART, `http://smart.embl.de`
49. Superfamily, `http://supfam.cs.bris.ac.uk`
50. TPRpred, `http://toolkit.tuebingen.mpg.de/tprpred`
51. Arabidopsis Genome Initiative (AGI, 2000), `http://www.arabidopsis.org/portals`
52. Pfam, `http://pfam.wustl.edu/hmmsearch.shtm`