BIOINFORMATICS ORIGINAL PAPER
Vol. 31 no. 3 2015, pages 332–339
doi:10.1093/bioinformatics/btu666

Sequence analysis
Advance Access publication October 9, 2014

Integrative data analysis indicates an intrinsic disordered domain character of Argonaute-binding motifs

Andrzej Zielezinski and Wojciech M. Karlowski*
Laboratory of Computational Genomics-Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, 61-614 Poznan, Poland

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Argonaute-interacting WG/GW proteins are characterized by the presence of repeated sequence motifs containing glycine (G) and tryptophan (W). The motifs seem to be remarkably adaptive to amino acid substitutions and their sequences show non-contiguity. Our previous approach to the detection of GW domains, based on scoring their gross amino acid composition, allowed annotation of several novel proteins involved in gene silencing. The accumulation of new experimental data and more advanced applications revealed some deficiency of the algorithm in prediction selectivity. Additionally, W-motifs, though critical in gene regulation, have not yet been annotated in any available online resources.

Results: We present an improved set of computational tools allowing efficient management and annotation of W-based motifs involved in gene silencing. The new prediction algorithms provide novel functionalities by annotation of the W-containing domains at the local sequence motif level rather than by overall compositional properties. This approach represents a significant improvement over the previous method in terms of prediction sensitivity and selectivity. Application of the algorithm allowed annotation of a comprehensive list of putative Argonaute-interacting proteins across eukaryotes. An in-depth characterization of the domains' properties indicates its intrinsic disordered character. In addition, we created a knowledge-based portal (whub) that provides access to tools and information on RNAi-related tryptophan-containing motifs.

Availability and implementation: The web portal and tools are freely available at http://www.comgen.pl/whub.

Contact: wmk@amu.edu.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 7, 2014; revised on September 11, 2014; accepted on October 6, 2014

1 INTRODUCTION

Despite the fact that different classes of small RNAs are generated by largely different biogenesis pathways, in order to function they associate with Argonaute proteins (AGOs) and form the RNA-induced silencing complex (RISC) (Kuhn and Joshua-Tor, 2013). The 'guide strand' of the sRNA duplex directs the AGOs to specific targets on the DNA (transcriptional gene silencing) or RNA (post-transcription gene silencing) through base complementarity. On the target molecule, the AGO proteins interact with GW proteins, which are characterized by the presence of repeated sequence motifs containing two amino acids: glycine (G) and tryptophan (W). These motifs mediate direct interaction with various AGO protein family members throughout the eukaryotic kingdom (Bednenko *et al.*, 2009; Ding *et al.*, 2005; Eulalio *et al.*, 2009b; He *et al.*, 2009; Jakymiw *et al.*, 2005; Karlowski *et al.*, 2010. Partridge *et al.*, 2007; Till *et al.*, 2007). Most recently, it has been shown that some viruses contain polypeptides that mimic host-encoded proteins with WG/GW motifs and act as bait for the Argonaute to hack into the host's essential effector of RNA silencing (Aqil *et al.*, 2013; Azevedo *et al.*, 2010; de Ronde *et al.*, 2014; Giner *et al.*, 2010; Szabó *et al.*, 2012).

Over the years, our view of GW proteins has evolved from them being a static platform for clustering of AGO proteins to functional domains composed of short tryptophan-containing modules that coordinate vital steps in gene silencing (Azevedo *et al.*, 2011). The extremely divergent sequence, variable lengths (ranging from 22 to up to 650 aa) and number of composing dipeptide WG/GW repeats (from 1 to up to 40) make the WG/GW domain difficult to identify using in silico methods (Karlowski *et al.*, 2010). Therefore, our previous approach (Karlowski *et al.*, 2010; Zielezinski and Karlowski, 2011) was based on scoring gross compositional amino acid properties of the sequence rather than plain linear similarity. However, the more advanced and detailed applications revealed some deficiency of the algorithm in prediction selectivity. Additionally, the initial method did not include identification of the recently described, new types of W-based motifs (referred to here as W-motifs) which are involved in interactions with the CCR4-NOT complex and mRNA translational repression and deadenylation processes (Chekulaeva *et al.*, 2011).

Deficiency of a comprehensive resource for functional GW/GW proteins in any of the central protein [e.g. UniProt (The UniProt Consortium, 2014), RefSeq (Pruitt *et al.*, 2012)] and/or specialized domain databases [e.g. Pfam (Finn *et al.*, 2014), InterPro (Hunter *et al.*, 2012)] as well as quickly accumulating knowledge about the properties of the W-based AGO/CCR4-NOT interacting proteins inspired us to develop new computational tools to facilitate efficient management and annotation of W-motifs. The novel prediction approach addresses all the faults of the previous method and provides novel functionalities by annotation of the tryptophan-containing domains involved in gene silencing at the local sequence motif level, rather than by searching for large regions of multiple WG/GW motifs. We developed Whub—a knowledge-based portal that provides

*To whom correspondence should be addressed.

access to all available information on RNAi-related W-containing motifs—in order to address the need for a centralized source of information about this structurally and functionally unusual group of proteins.

2 MATERIALS AND METHODS

2.1 Calculation of the position-specific scoring matrix for single W-containing motifs.

The source sequence dataset consists of a manually selected collection of WG/GW protein domains that have already been described as interacting with AGO/CCR4-NOT protein complexes. From the initial set all non-overlapping subsequences were extracted that contained a single W residue flanked by non-tryptophan amino acids. The ungapped multiple sequence alignment was performed to represent a profile of the selected motifs, in which tryptophan was assumed to be a midpoint (position 0) and the flanking non-W residues spread through positions in both directions (N- and C-termini). As the subsequences differ in their lengths, the weights of the W-flanking residues decrease. The observed frequencies (P_{obs}) of all non-tryptophan residues were obtained from counts of amino acids within each column of the profile following the formula:

$$\mathrm{P}_{\text{obsia}} = N_{\text{ia}}/N$$

where i stands for each of the positions in the motif sequence; a stands for each of the amino acids present in given position i; N_{ia} is the number of occurrences of amino acid a at given position i; N is the number of motif sequences. The observed frequencies were compared with the corresponding expected frequencies (P_{exp}) obtained from background W-motif sequences in UniProt and were used to calculate the log-odds according to the following formula: $D_{\text{ia}} = 2 \times \log2(P_{\text{obs}}/P_{\text{exp}})$. The method that was used for annotation of the W-containing motifs procedure is presented in Supplementary Figure S1.

2.2 Search for potential AGO-binding domains

Eukaryotic proteins retrieved from the UniProt database were scanned for potential AGO-binding sites using the position-specific scoring matrix (PSSM) built on published, experimentally verified W-motifs. *P*-values were calculated separately for W-motifs and assembled domains, and score values were used to perform distribution-fitting analysis using the EasyFit program (Mathwave Technologies). The proteins were clustered into families using the cd-hit program (Li and Godzik, 2006).

2.3 Benchmarking WG/GW identification methods

The prediction methods (PSI-BLAST, HMMER, Agos, Wsearch) were trained on a set of 200 orthologous AGO-binding protein sequences and 1000 series of randomly selected sequences from yeast, Arabidopsis and human proteomes. The prediction was evaluated by the 10-fold cross-validation approach. The whole dataset was randomly partitioned into 10 groups of approximately equal size. The methods were trained on nine groups and tested on the remaining set to ensure that the training process was completely independent of the test data. Each of the groups was used for benchmark calculation. The predicted quality results were evaluated by sensitivity (SN), specificity (SP) and precision (PPV) using the following formulas: $\mathrm{SN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, $\mathrm{SP} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$, $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, where TP, FN, FP and TN represent the numbers of true positive, false negative, false positive and true negative residues in the prediction, respectively.

2.4 Random forest-based classification of AGO-binding W-motifs

The AGO-binding activity was formulated as a binary classification problem: AGO-binding W-motifs are labeled as 1 and non-binding ones as 0. The sliding window technique was used to encode the amino acid residues flanking tryptophan. Each amino acid in the W neighborhood profile is characterized by several descriptors including the hydrophobic index (Kyte and Doolittle, 1982), flexibility parameters (Vihinen *et al.*, 1994), hydrophobicity (Hopp and Woods, 1981), relative accessible surface area (Chothia, 1976), amino acid weights and volumes. In addition, the distance (as the number of amino acids) from the central tryptophan to the closest W residue from N- and C-termini is used. The parameters selected to determine whether the W residue belongs to the AGO-binding class include its neighboring amino acid context and distances to the nearest W. Various window sizes were tested (Supplementary Table S1) and a value of 21 was selected for further applications. The Random Forest algorithm (Breiman, 2001) was implemented by the Scikit-learn machine learning python library (Pedregosa *et al.*, 2011).

3 RESULTS

3.1 Wsearch: the position-based prediction method for functional W-motifs annotation

The new algorithm (Wsearch) determines the probability that any given amino acid will be found in a particular position of an AGO-binding site based on an analysis of experimentally verified motifs from plants, animals and other eukaryotes. Wsearch allows for annotation of single W-containing motifs and identification of their boundaries as well as statistical qualification of predicted sequences.

The prediction procedure incorporates the concept of the PSSM, where each amino acid is scored based on its frequency in a given location within the motif. The rationale behind this approach comes from the experimental observation that tryptophan is critical for the domain function and its amino acid context modulates the strength of interaction. It has also been reported that not all WG/GW repeats located in the functional domain contribute equally to the interaction (Chekulaeva *et al.*, 2011; Eulalio *et al.*, 2009a; Takimoto *et al.*, 2009; Yao *et al.*, 2011), and that the substitution of residues adjacent to the repeats affects AGO binding (Pfaff *et al.*, 2013; Till *et al.*, 2007). Therefore, W is located in the midpoint of the calculated scoring table and the likelihoods of all other amino acids are spread in both directions. As a consequence of the variable W-motif sequence length (from 4 up to 95 amino acids), the residues located at more distant positions from the central tryptophan have a weaker impact on the motifs' scores than residues from closer surroundings (see ‘Materials and Methods’). This way the output of the prediction procedure includes the smallest functional units of the whole domain, which allows more detailed comparative analyses and should be better suited for subsequent experimental applications. Because the properties of the W-based, AGO-binding domains are still not well defined, during the prediction procedure we employ a minimum of assumptions strategy: all W-containing motifs in an analyzed sequence are treated as seeds for motif prediction. The annotation process runs in both directions from the tryptophan residue, calculating a cumulative score using values from the PSSM matrix. The analysis stops when the score reaches a maximum, which guarantees identification of the best-scoring motif.

3.1.1 Benchmarking of AGO-binding identification methods For the new method evaluation, we compared the quality of Wsearch predictions with a previous version of the algorithm (Agos) as well as two of the most popular domain identification approaches—PSI-BLAST (Altschul *et al.*, 1997) and HMMER3 (Finn *et al.*, 2011). The quality of the predictions was assessed on a domain and amino acid level.

3.1.2 Performance at the domain level All methods were trained and tested on a set of 200 orthologous AGO-binding proteins and 1000 series of 2000 randomly selected sequences from yeast, Arabidopsis and human proteomes. The results of a 10-fold cross-validation test (see ‘Materials and Methods’) show that Wsearch achieved best performance in identifying the highest number of AGO-binding proteins (96.9%) with the highest precision (99.5%). The Agos tool achieved the second best score with 90.9% sensitivity and 95.8% precision, while the sequence alignment-based tools reached values below 60%.

3.1.3 Performance at the amino acid level Recently, Pfaff *et al.* (2013) published results on the characterization of 20aa-long tryptophan-containing peptides from the human N-terminal part of the TNRC6B protein (162-996 aa) showing variable affinity for binding AGOs (Pfaff *et al.*, 2013). Figure 1 shows a schematic representation of the computational identification of Ago-binding motifs by Agos and Wsearch along with the results published by Pfaff *et al.*

Application of the Agos method resulted in the identification of a single continuous sequence region 616–903 covering 60% of amino acids (with 79% precision) defined by the peptide-screening experiment as AGO binding. In contrast, all nine AGO-interacting regions overlap with the Wsearch prediction results, including region 467–501 that does not contain the canonical WG/GW repeat. In total, the Wsearch method predicted 80% of the amino acids involved in AGO binding with 88% precision. Moreover, the new algorithm detected two additional fragments in this region which were not predicted by the laboratory approach but exhibited potential for binding AGO proteins. The quality of these predictions ($P = 1.61E\text{-}05$ and $P = 5.78E\text{-}04$) suggests that they could possibly interact with AGOs; however, the interaction may not be AGO2 specific. Several W-motifs present in the N-terminal part of the TNRC6B protein that have not been classified as AGO-interacting additionally confirmed high selectivity of the new motif prediction approach even though the whole N-terminus (162–996) was used to train the algorithm.

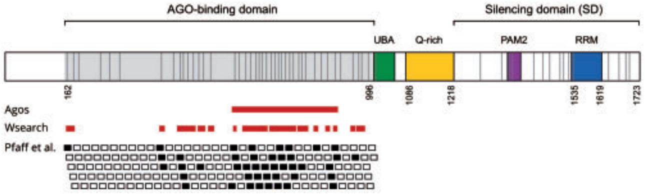


Fig. 1. Schematic representation of the TNRC6B domain architecture along with the results of computational annotation of AGO-binding motifs using Agos and Wsearch. The results of the peptide-screening experiment by Pfaff *et al.* (2013) are shown in the form of tiles of rectangles where the AGO-binding activity is indicated as black-filled squares

Additionally, we tested the performance of Agos and Wsearch on a set of 79 short peptides (20 aa) as designed by Pfaff *et al.* (2013) and analyzed by peptide array probing using the recombinant Ago2 protein (Supplementary Table S2). The Agos algorithm was unable to detect any of the AGO-binding motifs because the amino-acid composition signal coming from such short motifs was too weak to discriminate AGO-binding sites from background sequences. Wsearch correctly classified 60% of the AGO-binding peptides with 80% precision.

Compositionally similar reiterated W-motifs from the C-terminal region of the GW182 protein and its human homolog TNRC6A were reported by Lian *et al.* (2009) as being responsible for interaction with AGO2 proteins. In our predictions these sequences did not exhibit AGO-binding properties. This was confirmed by several other reports (Chekulaeva *et al.*, 2011; Eulalio *et al.*, 2009a; Fabian *et al.*, 2011; Lazzaretti *et al.*, 2009; Takimoto *et al.*, 2009; Zipprich *et al.*, 2009).

3.2 New properties of the Ago-binding domain

A careful bioinformatic examination of all the described, experimentally verified AGO-interacting motifs provides new insights into the domain’s properties.

The difference in the level of sequence conservation between plant and animal W-domains is one of the most striking features of AGO-binding motifs (Supplementary Figure S2). The sequences of plant domains are very variable and seem to be shaped by intense evolutionary activity. The results of all-versus-all pairwise alignments of W-containing domains in plants among the orthologs of NRPE1 and SPT5 proteins indicate a mean sequence identity of 21% and 29%, respectively (Supplementary Table S3). The other parts of the proteins show 55% and 46% identity, respectively. This suggests that AGO-interacting domains in plants evolve faster than other parts of the protein and diverge to a point where alignment-based similarity cannot be precisely determined even between closely related orthologs. In animals, however, AGO-binding domains seem to be more conserved and their sequences in GW182 or Prion proteins (PrP) retain high identity throughout the mammalian lineage (87–91%; Supplementary Table S3). Nevertheless, both plant and animal AGO-binding domains display no sequence conservation outside of the core tryptophan pattern, between proteins from different families (e.g. AGO-binding domains between NRPE1 and SPT5 or between the GW182 and PrP).

As was previously reported (Karlowski *et al.*, 2010), computer simulations of domain-swapping experiments between plant and animal AGO-binding proteins from different unrelated families revealed their evolutionary conservation of the amino acid composition. The tryptophan sequence surroundings in the AGO-interacting W-motifs of plants and animals show a biased but common ($P = 0.21$) amino-acid composition that deviates significantly ($P < 0.001$) from nonfunctional W-motifs and indicates a strong positive tendency towards small, polar and non-hydrophobic amino acids (G, N, D, E, S). The same tendency was observed based on position-specific amino acid preference analysis (Fig. 2; Supplementary Figure S3).

In general, it seems that in different AGO-binding proteins from plants and animals the tryptophan residues are embedded in hyper-variable low complexity sequences that fall into locally disordered regions with low overall hydropathy and high net charge.

An analysis of distance distribution between 6799 tryptophan residues from experimentally verified plant and animal AGO-binding domains indicates that in tandemly repeated motifs, tryptophans are in most cases separated by 8–13 amino acid residues (Fig. 3a). This means that each W in the functional motif is symmetrically surrounded by 4–7 amino acids on each side.

Recently, Schirle and MacRae (2012) resolved the structure of a complex of human AGO2 with two molecules of tryptophan bound to the site pockets of the PIWI domain (Schirle and MacRae, 2012). A distance of about 24 Å between the two tryptophan-binding sites, measured along the surface of the AGO2 molecule (Schirle and MacRae, 2012), corresponds to the span of 8–14 residues. Additionally, using biochemical and nuclear magnetic resonance (NMR) experiments, Pfaff *et al.* (2013) reported that two tryptophans from the TNRC6B protein are separated by a spacer with a minimal length of 10 aa. A similar mode of molecular recognition was also observed for the interaction of GW182 proteins with the CNOT9 protein (Chen *et al.*, 2014; Mathys *et al.*, 2014). Our comparative analysis of all functional AGO-interacting motifs complies well with the experimental results and additionally suggests that the distance between tryptophan residues in AGO-interacting domains is not restricted to the human GW182 protein but is globally conserved throughout eukaryotes.

In several hundred of the analyzed W-motifs the distance separating the closest, neighboring W residues spanned 8–14 amino acids, reaching in extreme cases 100 residues. The graphical presentation of positive and negative selection based on the PSSM scores that operate on each position relative to the W residue is shown in Figure 3b. Although the sum of positive log-odds decreases rapidly by one half at positions outside the ± 6 range, the amino acids at motif positions of –7, –8 and –10 as well as + 7, + 8 and + 9 still show clear signatures of positive selection. In contrast, further positions from the W midpoint show a lack of any positive, detectable amino acid preference. This result confirms the functional size of the AGO-binding motif that does not seem to exceed the ±10 amino acid boundary. Within this range, the positions of the strongest negative selection include locations ±1 and –2, within the closest proximity to the tryptophan residue. Such distribution of the conservative positions may imply a non-symmetric character of the motif. Taking into account the optimal size of the motif and the values from the PSSM array calculated using eukaryotic AGO-binding domains, the sequence of the most ideal, preferred single AGO-binding sequence is `ngnnnns nsgwgeppnqnss`. However, such a sequence motif has not been found in any of the known proteins.

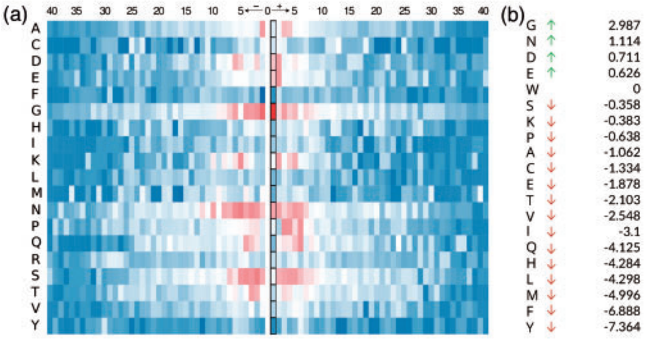


Fig. 2. Positional analysis of the amino-acid composition of 6799 W-motifs from experimentally verified AGO-binding proteins in Eukaryotes. (a) PSSM shown as a heat map; color blocks represent amino acid preferences that are present (red) or absent (blue) on a given motif position. (b) Example of amino acid preferences at position 1

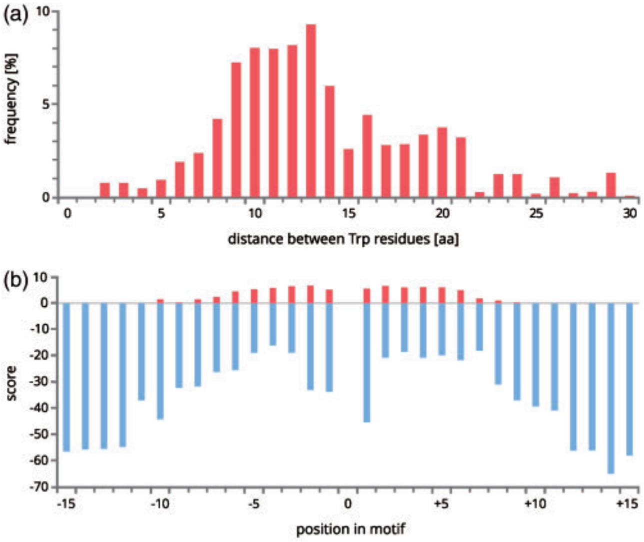


Fig. 3. W-motif length distribution of AGO-binding proteins in eukaryotes. (a) Sequence distances (amino acids) between tryptophan residues. (b) Graphical presentation of positive and negative selection at each position relative to the W residue as calculated in PSSM matrices. Bars represent the sum of the positive and negative values

Based on structural analyses determined by NMR (Zahn *et al.*, 2000) and X-ray (Knaus *et al.*, 2001), human PrP AGO-binding W-motifs are embedded within the unstructured N-terminal region. Furthermore, secondary structure predictions of the GW182 protein family suggest that the AGO-binding domain is unstructured and does not represent an independent folding unit (Eulalio *et al.*, 2009a). These studies, along with the results of the sequence properties as described above, suggest that the AGO-interaction motifs of eukaryotic proteins constitute a family of Intrinsically Disordered Domains (IDD) that exist as ensembles of rapidly interconverting conformations. The AGO-binding IDD domain can be characterized by the following distinct features: (i) AGO-binding motifs, which like many other disordered regions evolve faster than their structurally ordered regions, thus resulting in sequence divergence where positional similarity cannot be precisely determined; (ii) the biased amino acid composition of AGO-binding motifs is consistent with the low overall hydrophobicity and high net charge characteristic of IDDs. This is manifested by the exclusion of order-promoting amino acids with the exception of tryptophan (Y, F, I, V, L) rather than a positive preference for others (Figs. 2 and 3b); e.g. the AGO-binding sites are characterized by a very low content of cysteine residues (C), which is known to have a significant contribution to protein conformation stability via the disulfide bond formation or by being involved in the coordination of different prosthetic groups; (iii) the W-motif regions, just as many IDDs (Tompa, 2003), show excessive length polymorphism due to variation in the number and size of repeats (Fig. 3b). It was proposed that IDD domains evolve by repeat expansion (Tompa, 2003). In a similar way, AGO-binding domains fall into a continuum of repeat units where the exact number of W-motifs differs between highly related proteins, i.e. NRPE1 orthologs in various plants species or paralogs of human GW182 proteins.

The comparative analysis presented here suggests that AGO-binding domains, though significantly variable in sequence and length, are composed of repeated sequence motifs that span from 10 to 20 amino acids, placing the W residue at the center of the hydrophilic and charged surface. This may indicate that W-based, AGO-binding domains are constrained within a narrow subset of possible sequences, which most likely are the result of the biophysical restraints of AGO-interactions that have yet to be elucidated.

3.3 Identification of novel AGO-binding proteins

We applied Wsearch to computationally screen novel AGO-binding sites across eukaryotic proteomes. Table 1 shows a ranking list of the top-scoring protein families ($P \leq 1e\text{-}05$) with corresponding information about the taxonomic span and number of members (the full data are available as Supplementary Table S5).

Several top-scoring predictions overlap with previous results and in some cases correspond to well known and already characterized AGO-interacting proteins (e.g. SPT5, NRPE1 and GW182). Among the several novel AGO-interacting proteins, DEAD-box RNA helicases and heterogeneous nuclear ribonucleoproteins (hnRNPs) draw special attention. The well-conserved helicase domain of the DEAD-box proteins of fungi (76), animals (18), plants (21) and protists (5) is adjacent or flanked by highly variable tryptophan-containing repeat sequences (data not shown). These proteins were reported to localize to P-bodies in yeast (Beckham *et al.*, 2008), were found in human RISC complexes and, in addition to other functions, are implicated in translation initiation, translation repression and RNAi (Chen *et al.*, 2014; Mathys *et al.*, 2014; Parsyan *et al.*, 2011). All identified hnRNP proteins are characterized by the presence of one or two well-conserved RNA-Recognition Motifs (RRM) and a highly divergent C-terminal region that is rich in tryptophan-containing peptides (Supplementary Figure S4). The top-ranking candidates, identified in Nematoda and Arthropoda, are already implicated in gene silencing: the A1 protein from *C. elegans* has been reported to bind the conserved loop of pri-miR-18a through RRM domains (Michlewski *et al.*, 2008), and two other proteins (Hrb98DE and Hrb87F) from the fruit fly co-precipitate with the CCR4-NOT complex (Temme *et al.*, 2010).

The other predicted AGO-interacting proteins listed in Table 1 are represented by peptides of narrow phylogenetic distribution and are mainly located in unrelated sequence families; e.g. the IF-2 Translation Initiation Factor with 70 tryptophan motifs can be found only in two species from the genus *Cryptococcus*: *C. gatti* and *C. neoformans*. Another case of such non-conserved sequences includes two RNA-binding proteins from *Paramecium tetraurelia*: Nowa1p and Nowa2p, which were previously suggested to bind AGO (Nowacki *et al.*, 2005).

3.4 Whub—a knowledge-based portal for silencing-related tryptophan-motifs

W-motifs, though shown to be critical in post-transcriptional gene regulation, have not yet been annotated in any available online resources. Moreover, it is unlikely that a researcher will find a complete list of currently known AGO-binding proteins across central databases and/or protein specialized resources. In addition, going through the literature to gain a general view of functional W-containing motifs is impractical, especially for newcomers, as each expert research group often focuses on a single species-specific protein. One of the outcomes of our study on the properties of AGO-interacting proteins is an extensive, high-quality collection of all publicly available literature and sequence data. Therefore, for present and future applications we created an integrated web portal (Whub; http://www.comgen.pl/whub) to facilitate efficient management of information about proteins containing W-motifs involved in gene silencing. Whub offers integrated and user-friendly access to current information about experimentally verified AGO-interacting proteins.

The portal offers three ways to access the data: the protein portlet, the bibliography browser and the prediction tools. The main protein entry section provides a visual map of tryptophan locations in the context of a protein's functional regions. This part of the portal allows to browse experimentally mapped, minimal W-rich regions that are sufficient for interaction with different AGO/CCR4-NOT proteins as well as a list of experimentally evaluated mutations. The information contained in the experimental results panel is graphically referenced to a full-length amino acid sequence. The foundation for the portal comes from information extracted by literature mining. The whole collection of papers describing AGO/CCR4-NOT interacting proteins is assembled in a graphical browser that allows to search for and sort the entries. The publication section is deeply integrated with other parts of the portal and creates a unique way of exploring the content through bibliographical data.

All of the information contained in the database section of the Whub portal represents the most up-to-date foundation for the W-motif annotation tool (Wsearch). The predictions

Table 1. List of predicted eukaryotic AGO-binding proteins

GW protein family						Representative protein	
Taxonomy range[a]	Organisms	Name	Short Description	Members	AGO-binding	UniProt AC	Total score
Eukaryota	38	SPT6	Transcription elongation factor Spt6	58	candidate [1]	B4P508	486.97
Eukaryota	122	SPT5	Transcription elongation factor Spt5	171	AGO4 [2,3]	M4CZ23	404.9
Eukaryota	84	DEAD/DEAH box	DEAD/DEAH box helicase family	120	NT	A2D755	353.62
Oikopleura	1	—	Uncharacterized	2	NT	E4Z1X2	318.7
Metazoa	94	GW182	Trinucleotide repeat-containing gene	285	AGO1-4 [4]	M7B672	279.41
Polysphondylium	1	—	Uncharacterized	1	NT	D3BJ75	262.37
Eukaryota	2	xylA-like	Bifunctional endo-1,4-beta-xylanase xylA-related	2	NT	A2FUM7	258.85
Glycine	1	—	Uncharacterized	1	NT	I1NFI4	252.82
Glarea	1	—	Uncharacterized protein	1	NT	S3DME1	195.3
Tetrahymena	1	CnjB	Zinc knuckle family	2	Twi1p [5]	Q24BQ3	194.4
Sclerotiniaceae	2	—	Uncharacterized	3	NT	A7EU40	183.79
Eukaryota	32	GRP	Uncharacterized glycine rich proteins	62	NT	B4GYA0	181.21
Nannochloropsis	1	PAP2	Pap2 haloperoxidase domain-containing protein	1	NT	K8YVY0	172.52
Caenorhabditis	5	PQN	Prion-like-(Q/N-rich)- domain-bearing protein	11	NT	A8XSU8	171.48
Paramecium	1	NOWA	NOWA1, NOWA2	3	candidate [8]	A0CDB6	161.8
Filobasidiella	2	IF2	Translation initiation factor IF-2	4	NT	J9VH01	154.66
Trichoplax	1	—	Predicted protein	1	NT	B3S6S9	154.27
Tetrahymena	1	WAG1	GW repeat protein	2	Twi1p [5]	B8XQC5	148.77
Magnoliophyta	13	NRPE1	Polymerase V subunit	19	AGO4 [6]	Q5D868	142.18
Oxytricha	1	Nucleoporin	Nucleoporin	1	NT	J9J4X8	132.44
Magnaporthaceae	2	—	Uncharacterized	2	NT	M4G6T7	127.85
Musca	1	—	Uncharacterized	1	NT	T1PKW5	125.96
Vitis	1	—	Uncharacterized	1	NT	F6HND0	124.29
Agaricomycotina	8	RNA_pol	DNA-directed RNA polymerase	11	NT	S8FH34	122.47
Marssonina	1	—	Uncharacterized	1	NT	K1WUC6	117.62
Eukaryota	58	hnRNP/RRM	Heterogeneous nuclear ribonucleoprotein	109	NT	Q4Q4J4	117.48
Ichthyophthirius	1	—	Universal minicircle sequence binding protein, putative	1	NT	G0QMY8	114.7
Panicoideae	2	—	Uncharacterized protein	2	NT	K3Z3H0	110.74
Coprinopsis	1	—	Putative uncharacterized protein	1	NT	A8N305	110.12
Mammalia	152	PrP	PrP	401	AGO1,2 [7]	Q16409	105.12

[a] highest taxon encompasses predicted AGO-binding proteins and their orthologs [1] (Karlowski *et al*., 2010), [2] (He *et al*., 2009), [3] (Bies-Etheve *et al*., 2009), [4] (Eulalio *et al*., 2009b), [5] (Bednenko *et al*., 2009), [6] (El-Shami *et al*., 2007), [7] (Gibbings *et al*., 2012), [8] (Nowacki *et al*., 2005).

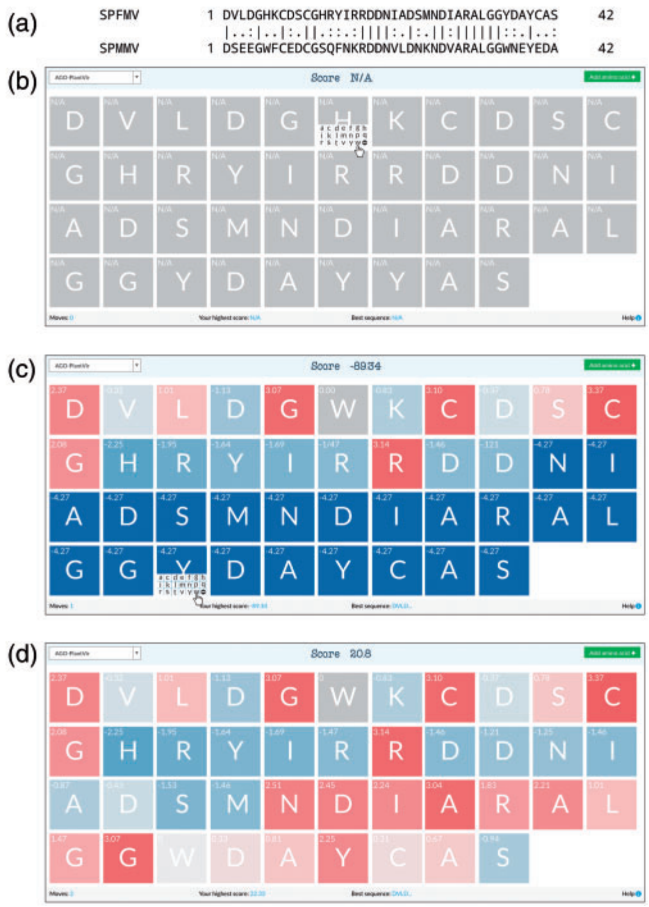
are shown in the form of interactive tables, charts and graphs. For large-scale, high-throughput sequence analysis applications, a stand-alone version of the prediction tool is provided which additionally allows for calculation of custom PSSM profiles.

The tool repertoire for AGO-binding site annotation is supplemented with a machine learning (Random Forest; RF) implementation of the prediction algorithm. In this approach each amino acid residue is characterized by a number of descriptors, including properties of amino acids and distances to the nearest tryptophan residue (for details see ‘Material and Methods’). The RF-based classifier achieves 95.17% sensitivity, 99.94% specificity and 99.68% precision. All evaluation test results are provided in Supplementary Table S1.

3.4.1 Computationally-aided domain design Recently, several research groups studied the detailed properties of the AGO- and CCR4-NOT-binding domains by biochemical mutagenesis assays. Szabó *et al*. (2012) transformed the nonfunctional part of the SPFMV P1 protein into a functional silencing suppressor with AGO-binding capacity by introducing two additional WG/GW motifs. In another study, Filipowicz’s group showed that insertion of at least four tryptophan residues into the unstructured region of the non-RNAi-related yeast protein Sic1p21 was sufficient to induce mRNA repression by recruiting the CCR4-NOT complex (Chekulaeva *et al*., 2011).

To model the substitution and insertion effects of various amino acids within the framework of presented properties of AGO/CCR4-NOT-binding motifs, we created an interactive tool that allows dynamic modification of any W-motif with a single residue resolution providing real-time, color-encoded prediction of binding quality (Fig. 4).

This puzzle-like interface represents a framework for simulation of mutagenesis experiments and hypothesis testing. As a case example, Figure 4 shows a simulation of Szabo’s experiment (Szabó *et al*., 2012)—rendering the functional AGO-binding domain from the nonfunctional SPFMV P1 ortholog. The amino acid sequences are presented as a string of blocks of



```
SPFMV     1 DVLDGHKCDSCGHRYIRRDDNIADSMNDIARALGGYDAYCAS   42
            |..:|..|:.||.:::.:||||:.|:.||:||||||::.|..:
SPMMV     1 DSEEGWFCEDCGSQFNKRDDNVLDNKNDVARALGGWNEYEDA   42
```

Fig. 4. Reconstruction of the Szabó *et al.* (2012) experiment with the Whub domain design tool. (a) Pairwise sequence alignment of homologous P1 protein fragments from SPFMV (non-AGO-binding) and SPMMV (AGO-binding) viruses. (b) Initial screen of the domain design interface showing no AGO-binding properties of P1 in SPFMV due to the lack of tryptophan. (c) Result of first substitution (H > W) in position 6 (indicated by the cursor on B). (d) Final screen after second substitution (Y > W) in position 36 (indicated by the cursor on C)

different intensity of red and blue colors reflecting PSSM scoring. In a more general gaming context one can consider the goal of the framework to have the ability to either modify the sequence in order to find the maximum scoring motif or to forecast the potential effect of future mutagenesis studies.

4 DISCUSSION

The most important aspect of today's digital biology is access to data. The increasing amount of information requires new tools and approaches that will allow for fast and accurate information retrieval and data mining. Very often the data are spread across many databases and publications. This is evident especially in the case of new and developing subjects of scientific research that have not yet been completely defined and have not been enclosed by mature tools and resources. Research on WG/GW proteins represents one example of such research enterprises—the proteins are investigated on local bases in which scientists concentrate on single proteins from particular organisms. However, the available accumulated knowledge would allow for a more global and synthetic view of this very interesting group of proteins.

The Whub portal is a user-oriented and comprehensive approach for collecting all available knowledge on the biology of AGO/CCR4-NOT-interacting W-motifs. We have created an efficient and attractive interface that facilities exploration of up-to-date information, the discovery of new relationships between data records and drawing biologically relevant conclusions concerning highly specific W-containing motifs.

The implementation of new computational tools involving a position-based approach instead of the composition-based prediction method resulted in the annotation of new groups of proteins that may interact with Argonautes. These include, among others, RRM-containing hnRNP proteins and DEAD-box helicases. Many promising predictions are already the subject of experimental verification. Our methods use the modern information technologies of machine learning and move the AGO-protein identification process to a new level, i.e. where the user receives a clear functional classification of a tested sequence instead of a set of score numbers that is difficult to interpret. The annotation results show that the presented computational approach can distinguish AGO-binding sites from other functional W-containing motifs (e.g. CCR4-NOT) despite the high amino-acid composition similarity between these two types of sequences (Supplementary Table S4). This supports the finding by Chekulaeva *et al.* (2011) about the distinct functions of these two W-based domains. Unfortunately, the limited number of available CCR4-NOT-binding motifs did not allow to build prediction models for this group of sequences. Nevertheless, we have already implemented all of the functionality into the Whub portal, but the quality of the predictions may vary. As soon as new data are available, they will be incorporated into the prediction procedure.

An in-depth analysis of the properties of experimentally verified and novel predicted proteins allowed for classification of the W-based domains into the IDD class. A fast evolution rate, biased amino acid composition, single amino-acid repetitions (Lobanov *et al.*, 2014) and excessive length polymorphism conform well to the characteristics of IDDs. This expands knowledge on this new and intensively investigated group of functional domains. On the other hand, it may help in further research on AGO/CCR4-NOT-interaction domains including computational and experimental approaches.

ACKNOWLEDGEMENT

A.Z. was a scholarship holder within the project 'Scholarship support for PhD students specializing in majors strategic for Wielkopolska's development', Sub-measure 8.2.2 Human Capital.

Funding: This work was supported by grants from the National Center of Science [UMO-2011/03/B/NZ2/01416 to W.M.K. and UMO-2011/03/N/NZ2/01440 to A.Z.] and KNOW RNA Research Centre in Poznan (No. 01/KNOW2/2014).

Conflict of interest: None declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Aqil,M. *et al.* (2013) The HIV-1 nef protein binds Argonaute-2 and functions as a viral suppressor of RNA interference. *PLoS One*, **8**, e74472.

Azevedo,J. *et al.* (2010) Argonaute quenching and global changes in Dicer homeostasis caused by a pathogen-encoded GW repeat protein. *Genes Dev.*, **24**, 904–915.

Azevedo,J. *et al.* (2011) Taking RISCs with Ago hookers. *Curr. Opin. Plant Biol.*, **14**, 594–600.

Beckham,C. *et al.* (2008) The DEAD-box RNA helicase Ded1p affects and accumulates in Saccharomyces cerevisiae P-bodies. *Mol. Biol. Cell*, **19**, 984–993.

Bednenko,J. *et al.* (2009) Two GW repeat proteins interact with *Tetrahymena thermophila* Argonaute and promote genome rearrangement. *Mol. Cell. Biol.*, **29**, 5020–530.

Bies-Etheve,N. *et al.* (2009) RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO Rep.*, **10**, 649–654.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Chekulaeva,M. *et al.* (2011) miRNA repression involves GW182-mediated recruitment of CCR4-NOT through conserved W-containing motifs. *Nat. Struct. Mol. Biol.*, **18**, 1218–1226.

Chen,Y. *et al.* (2014) A DDX6-CNOT1 complex and W-binding pockets in CNOT9 reveal direct links between miRNA target recognition and silencing. *Mol. Cell*, **54**, 737–750.

Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–12.

Ding,L. *et al.* (2005) The developmental timing regulator AIN-1 interacts with miRISCs and may target the Argonaute protein ALG-1 to cytoplasmic P bodies in C. elegans. *Mol. Cell*, **19**, 437–447.

El-Shami,M. *et al.* (2007) Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components. *Genes Dev.*, **21**, 2539–2544.

Eulalio,A. *et al.* (2009a) A C-terminal silencing domain in GW182 is essential for miRNA function. *RNA*, **15**, 1067–77.

Eulalio,A. *et al.* (2009b) The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. *RNA*, **15**, 1433–1442.

Fabian,M.R. *et al.* (2011) miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT. *Nat. Struct. Mol. Biol.*, **18**, 1211–1217.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Gibbings,D. *et al.* (2012) Human prion protein binds Argonaute and promotes accumulation of microRNA effector complexes. *Nat. Struct. Mol. Biol.*, **19**, 517–524, S1.

Giner,A. *et al.* (2010) Viral protein inhibits RISC activity by Argonaute binding through conserved WG/GW motifs. *PLoS Pathog.*, **6**, e1000996.

He,X.-J. *et al.* (2009) An effector of RNA-directed DNA methylation in Arabidopsis is an ARGONAUTE 4- and RNA-binding protein. *Cell*, **137**, 498–508.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl Acad. Sci. U. S. A.*, **78**, 3824–3828.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Jakymiw,A. *et al.* (2005) Disruption of GW bodies impairs mammalian RNA interference. *Nat. Cell Biol.*, **7**, 1167–1174.

Karlowski,W.M. *et al.* (2010) Genome-wide computational identification of WG/GW Argonaute-binding proteins in Arabidopsis. *Nucleic Acids Res.*, **38**, 4231–4245.

Knaus,K.J. *et al.* (2001) Crystal structure of the human prion protein reveals a mechanism for oligomerization. *Nat. Struct. Biol.*, **8**, 770–774.

Kuhn,C.-D. and Joshua-Tor,L. (2013) Eukaryotic Argonautes come into focus. *Trends Biochem. Sci.*, **38**, 263–271.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lazzaretti,D. *et al.* (2009) The C-terminal domains of human TNRC6A, TNRC6B, and TNRC6C silence bound transcripts independently of Argonaute proteins. *RNA*, **15**, 1059–1066.

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

Lian,S.L. *et al.* (2009) The C-terminal half of human Ago2 binds to multiple GW-rich regions of GW182 and requires GW182 to mediate silencing. *RNA*, **15**, 804–813.

Lobanov,M.Y. *et al.* (2014) HRaP: database of occurrence of HomoRepeats and patterns in proteomes. *Nucleic Acids Res.*, **42**, D273–D278.

Mathys,H. *et al.* (2014) Structural and biochemical insights to the role of the CCR4-NOT complex and DDX6 ATPase in microRNA repression. *Mol. Cell*, **54**, 751–765.

Michlewski,G. *et al.* (2008) Posttranscriptional regulation of miRNAs harboring conserved terminal loops. *Mol. Cell*, **32**, 383–393.

Nowacki,M. *et al.* (2005) Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in *Paramecium tetraurelia*. *Curr. Biol.*, **15**, 1616–1628.

Parsyan,A. *et al.* (2011) mRNA helicases: the tacticians of translational control. *Nat. Rev. Mol. Cell Biol.*, **12**, 235–245.

Partridge,J.F. *et al.* (2007) Functional separation of the requirements for establishment and maintenance of centromeric heterochromatin. *Mol. Cell*, **26**, 593–602.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pfaff,J. *et al.* (2013) Structural features of Argonaute-GW182 protein interactions. *Proc. Natl Acad. Sci. U. S. A.*, **110**, E3770–E3779.

Pruitt,K.D. *et al.* (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.

de Ronde,D. *et al.* (2014) Analysis of Tomato spotted wilt virus NSs protein indicates the importance of the N-terminal domain for avirulence and RNA silencing suppression. *Mol. Plant Pathol.*, **15**, 185–195.

Schirle,N.T. and MacRae,I.J. (2012) The crystal structure of human Argonaute2. *Science (80-.).*, **336**, 1037–1040.

Szabó,E.Z. *et al.* (2012) Switching on RNA silencing suppressor activity by restoring Argonaute binding to a viral protein. *J. Virol.*, **86**, 8324–8327.

Takimoto,K. *et al.* (2009) Mammalian GW182 contains multiple Argonaute-binding sites and functions in microRNA-mediated translational repression. *RNA*, **15**, 1078–1089.

Temme,C. *et al.* (2010) Subunits of the Drosophila CCR4-NOT complex and their roles in mRNA deadenylation. *RNA*, **16**, 1356–1370.

The UniProt Consortium (2014) UniProtKB/Swiss-Prot protein knowledgebase release 2014_04 statistics.

Till,S. *et al.* (2007) A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat. Struct. Mol. Biol.*, **14**, 897–903.

Tompa,P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *BioEssays*, **25**, 847–855.

Vihinen,M. *et al.* (1994) Accuracy of protein flexibility predictions. *Proteins*, **19**, 141–149.

Yao,B. *et al.* (2011) Divergent GW182 functional domains in the regulation of translational silencing. *Nucleic Acids Res.*, **39**, 2534–2547.

Zahn,R. *et al.* (2000) NMR solution structure of the human prion protein. *Proc. Natl Acad. Sci.*, **97**, 145–150.

Zielezinski,A. and Karlowski,W.M. (2011) Agos—a universal web tool for GW Argonaute-binding domain prediction. *Bioinformatics*, **27**, 1318–1319.

Zipprich,J.T. *et al.* (2009) Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. *RNA*, **15**, 781–793.