Beyond sequence similarity
– the curious case of GW/WG protein domain

WOJCIECH M. KARLOWSKI

Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznań, Poland

e-mail: wojciech.karlowski@amu.edu.pl

RNA silencing is an essential component of gene expression regulation. Small RNAs (20-30 nt) carry the sequence information that guides the silencing effector complexes to specific targets on DNA (Transcriptional Gene Silencing) or RNA (Posttranscriptional Gene Silencing) (Law and Jacobsen, 2010; Brodersen et al., 2008). Argonaute (AGO) proteins are factors common to all these complexes (Czech and Hannon, 2010), but their complete protein composition is not yet sufficiently explored. A recent advancement in this subject is the observation that a novel element of such complexes in eukaryotes contains an unusual, repetitive peptide sequence binding AGO proteins – the GW/WG motif (Partridge et al., 2007; Till et al., 2007; El-Shami et al., 2007).

The WG/GW motifs were first described in the human GW182 protein, which was postulated at that time to target specific transcripts to cytoplasmic GW bodies (Eystathioy et al., 2003). Since then, AGO-binding properties have been demonstrated for several WG/GW motif-containing proteins, including GW182 homologs in *Caenorhabditis elegans* (AIN-1 and AIN-2) (Zhang et al., 2007), Drosophila (GAWKY) (Behm-Ansmant et al., 2006), *Schizosaccharomyces pombe* (Tas3) (Partridge et al., 2007) and *Arabidopsis thaliana* large subunit of RNA PolV (NRPE1) (El-Shami et al., 2007).

Strikingly, a comparison of the *Arabidopsis* NRPE1 GW/WG sequence with sequences of other plants shows little conservation in the repeats other than the WG/GW pairs, even between relatively closely related species. Figure 1 shows a dot-plot, representing comparison of amino acid sequences for the large subunit of PolV from *Arabidopsis* and *Vitis vinifera*, where the whole sequence is well conserved with the exception of the region containing GW/GW domain (Fig. 1, indicated by red box). Despite low sequence similarity, the domains re-
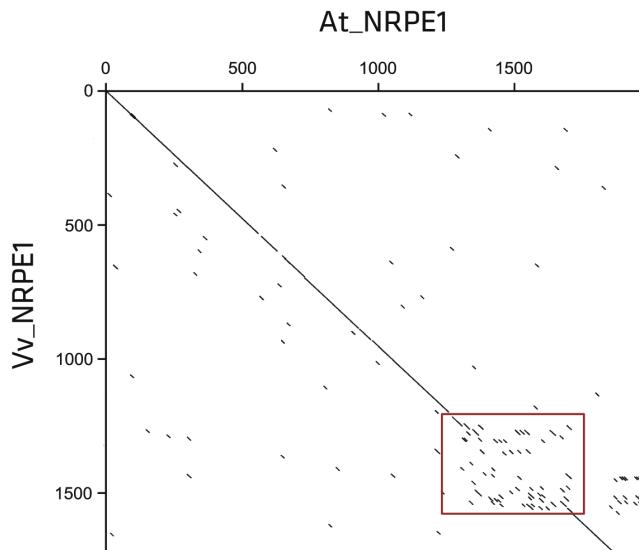
Fig. 1. Dot-plot comparison of amino acid sequences of PolV large subunits from *Arabidopsis thaliana* and *Vitis vinifera*. Red box marks repetitive WG/GW AGO-binding domain

veal a common amino acid composition, being rich in glycine, serine and tryptophan with low levels of cysteine, phenylalanine, histidine, methionine and tyrosine (Karlowski et al., 2010). The lack of sequence similarity, however, does not influence the functional universality of the domain. As demonstrated by El-Shami and coworkers (El-Shami et al., 2007) in a domain swapping experiment, the human sequence with GW/WG motifs of GW182 protein can substitute an analogous region in the *Arabidopsis* NRPE1 protein. Such chimeric construct binds to AGO4 *in vitro* and restores its DNA methylation activity, while the amino acid sequences show barely detectable similarity (El-Shami et al., 2007).

The unusual character of the AGO-binding WG/GW domain imposes several challenges and raises important questions. The highly divergent sequence, variable ami-

no acid lengths of the motifs and irregular number of repeats make the detection of the domain very difficult. Therefore, one of the most important tasks is to search for methods that will allow efficient annotation of this type of sequences. Most of the current annotation tools rely on sequence preservation – higher sequence similarity means a conserved function and a higher probability of correct protein description. However, in the case of WG/GW domain, the sequence shows a high level of divergence. To address this issue, we have developed in our laboratory a comprehensive set of bioinformatics tools which allow reliable identification of the domain, determination of its boundaries as well as statistical quantification of the predictions' quality (Zielezinski and Karlowski, 2011). Additionally, employing a method similar to the domain swapping experiment by El-Shami group (El-Shami et al., 2007), we have shown *in silico* that the new annotation algorithm, which is based on the sequence composition rather than the linear similarity, properly encapsulates the important features of the GW/GW motif. In our experiment, scoring tables calculated on plant or animal sequences were used to annotate animal or plant proteomes, respectively. In this way, we have shown that despite the non-detectable sequence similarity, the cross-taxonomic approach works well, and demonstrates conserved features of the functional AGO-binding domain. Moreover, the tool has been successfully applied to *Arabidopsis* whole genome annotation and the results survived laboratory validation (Karlowski et al., 2010). This approach expands the arsenal of annotation tools by allowing the identification of poorly conserved domains with AGO-binding proteins.

At the moment, we can annotate the AGO-binding domains in all organisms with high statistical confidence. However, several crucial questions still remain unanswered. One of the most important issues concerns the presence of other low conserved functional domains. Is their structure also based on the presence of the tryptophan residues? Which other amino acids in such domains are crucial and which can be easily substituted? Evidently, questions like this can only be answered by concurrent efforts of bioinformatics and laboratory approaches.

Another interesting aspect of the WG/GW domain properties concerns its unusual sequence diversity. It would be quite interesting to obtain a deeper understanding of the evolutionary mechanisms responsible for the shaping of the non-conserved character of the domain. Obviously, there are some differences in the degree of sequence preservation in plants and animals, which raise a very interesting question about the diversity of the evolutionary mechanism shaping the domain in various taxonomic groups. Recently, it has been shown that viruses encode proteins harboring the functional GW/WG motifs, which leads to the suppression of host silencing mechanisms (Bivalkar-Mehla et al., 2010). In this context, the unusual diversity of closely related sequences may be explained by a molecular race for efficiency of AGO-binding properties.

The lack of sequence conservation also imposes undetermined structural properties of the domain. It resembles, somehow, the case of recently described Intrinsically Disordered Domains (IDD) (Ward et al., 2004). Therefore, it is possible that besides conserved tryptophan and glycine residues in the GW/WG domain, other components are responsible for the creation of the binding environment, and dynamically adjust to the interaction properties of the complex.

Recently, research on the function of WG/GW domains has become even more complex and interesting. Three papers have been published (Braun et al., 2011; Chekulaeva et al., 2011; Fabian et al., 2011) that depict a novel function for the GW-based motifs. It has been demonstrated that C-terminal regions of the GW182 proteins from humans and *Drosophila* function by recruiting components of the CCR4–NOT complex inducing deadenylation and mRNA decay. Moreover, the Filipowicz group (Chekulaeva et al., 2011) have clearly presented that tryptophan residues located in the C-terminal region of the protein, similar as in AGO-binding activity, are required and sufficient for the interaction with silencing complex. In an additional and rather spectacular experiment, the authors introduced random mutations that transformed amino acids in the unstructured fragment of the yeast protein Sic1p21 into tryptophan. Interestingly, such artificially generated sequences, somewhat resembling the W-motifs, were able to induce mRNA repression by recruiting CCR4-NOT (Chekulaeva et al., 2011). These findings provoke new challenges for computational annotation algorithms, because at the current stage method developed in our laboratory can only classify the CCR4-NOT affinity as a low probable AGO binding domain. New information regarding the properties of the WG/GW domains and modification of statistical

methods is clearly required to allow efficient and reliable identification of this new function of the W-domain.

The newly discovered molecular roles for W-based motifs fill another gap in the description of functional small RNA silencing complexes. At the same time they raise new questions and open new exciting perspectives for creative research, where computational methods play a crucial role in exploring new solutions, seeking new answers and going where no scientist has gone before. It is clear that the quest for not-yet-identified proteins with AGO or CCR4-NOT binding domains has just begun. The question if, two identified functions saturate all the possible activities of W-based domains, remains open. However, it is obvious now that the time has come to focus on the non-conserved, unstructured parts of proteins with more scientific attention.

References

Behm-Ansmant I., Rehwinkel J., Doerks T., Stark A., Bork P., Izaurralde E. (2006) *mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes*. Genes Dev. 20: 1885-1898. 10.1101/gad.1424106

Bivalkar-Mehla S., Vakharia J., Mehla R., Abreha M., Kanwar J.R., Tikoo A., Chauhan A. (2011) *Viral RNA silencing suppressors (RSS): Novel strategy of viruses to ablate the host RNA interference (RNAi) defense system*. Virus Res. 155(1): 1-9. 10.1016/j.virusres.2010.10.003

Braun J.E., Huntzinger E., Fauser M., Izaurralde E. (2011) *GW182 Proteins Directly Recruit Cytoplasmic Deadenylase Complexes to miRNA Targets*. Mol. Cell 44: 120-133. 10.1016/j.molcel.2011.09.007

Brodersen P., Sakvarelidze-Achard L., Bruun-Rasmussen M., Dunoyer P., Yamamoto Y.Y., Sieburth L., Voinnet O. (2008) *Widespread Translational Inhibition by Plant miRNAs and siRNAs*. Science 320: 1185-190. 10.1126/science.1159151

Chekulaeva M., Mathys H., Zipprich J.T., Attig J., Colic M., Parker R., Filipowicz W. (2011) *miRNA repression involves GW182-mediated recruitment of CCR4-NOT through conserved W-containing motifs*. Nat. Struct. Mol. Biol. 18(11): 1218-1226. 10.1038/nsmb.2166.

Czech B., Hannon G.J. (2010) *Small RNA sorting: matchmaking for Argonautes*. Nat. Rev. Genet. 12: 19-31. 10.1038/nrg2916

El-Shami M., Pontier D., Lahmy S., Braun L., Picart C., Vega D., Hakimi M.A., Jacobsen S.E., Cooke R., Lagrange T. (2007) *Reiterated WG/GW motifs form functionally and evolutionarily conserved ARGONAUTE-binding platforms in RNAi-related components*. Genes Dev. 21: 2539-2544. 0.1101/gad.451207

Eystathioy T., Jakymiw A., Chan E.K.L., Séraphin,B., Cougot N., Fritzler M.J. (2003) *The GW182 protein colocalizes with mRNA degradation associated proteins hDcp1 and hLSm4 in cytoplasmic GW bodies*. RNA 9: 1171-1173.

Fabian M.R., Cieplak M.K., Frank F., Morita M., Green J., Srikumar T., Nagar B., Yamamoto T., Raught B., Duchaine T.F. et al. (2011) *miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4-NOT*. Nat. Struct. Mol. Biol. 18(11): 1211-1217. 10.1038/nsmb.2149.

Karlowski W.M., Zielezinski A., Carrère J., Pontier D., Lagrange T., Cooke R. (2010) *Genome-wide computational identification of WG/GW Argonaute-binding proteins in Arabidopsis*. Nucl. Acids Res. 38: 4231-4245. 10.1093/nar/gkq162

Law J.A., Jacobsen S.E. (2010) *Establishing, maintaining and modifying DNA methylation patterns in plants and animals*. Nat. Rev. Genet. 11: 204-220. 10.1038/nrg2719

Partridge J.F., DeBeauchamp J.L., Kosinski A.M., Ulrich D.L., Hadler M.J., Noffsinger V.J.P. (2007) *Functional separation of the requirements for establishment and maintenance of centromeric heterochromatin*. Mol. Cell. 26: 593-602. 10.1016/j.molcel.2007.05.004

Till S., Lejeune E., Thermann R., Bortfeld M., Hothorn M., Enderle D., Heinrich C., Hentze M.W., Ladurner A.G. (2007) *A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain*. Nat. Struct. Mol. Biol. 14: 897-903. 10.1038/nsmb1302

Ward J.J., Sodhi J.S., McGuffin L.J., Buxton B.F., Jones D.T. (2004) *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life*. J. Mol. Biol. 337: 635-645. 10.1016/j.jmb.2004.02.002

Zhang L., Ding L., Cheung T.H., Dong M.-Q., Chen J., Sewell A.K., Liu X., Yates J.R., Han M. (2007) *Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2*. Mol. Cell 28: 598-613. 10.1016/j.molcel.2007.09.014.

Zielezinski A., Karlowski W.M. (2011) *Agos – a universal web tool for GW Argonaute-binding domain prediction*. Bioinformatics 27(9): 1318-1319. 10.1093/bioinformatics/btr128