BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 9 2011, pages 1318–1319
doi:10.1093/bioinformatics/btr128

Sequence analysis

Advance Access publication March 8, 2011

Agos—a universal web tool for GW Argonaute-binding domain prediction

Andrzej Zielezinski and Wojciech M. Karlowski*

Laboratory of Computational Genomics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, 61-614 Poznan, Poland

Associate Editor: John Quackenbush

ABSTRACT

Motivation: AGO(Argonaute)-binding domains, composed of repeated motifs, in which only binary combinations of tryptophan and glycine are conserved, bind AGO proteins and are essential during RNAi-mediated gene silencing. The amino acid sequence of this domain is extremely divergent and therefore very difficult to detect. Commonly used bioinformatic tools fail to identify tryptophan–glycine and/or glycine–tryptophan motifs (WG/GW) domains and currently there is no publicly available software which can detect these weakly conserved, but functional AGO-binding segments.

Results: Recently, we have developed an algorithm based on compositional analysis of the amino acid content of the domain. We have demonstrated that the algorithm can be successfully applied for the identification of the new WG/GW proteins in the *Arabidopsis* genome. Here we introduce Agos (Argonaute-binding domain screener), a novel universal web service for *de novo* identification of WG/GW domains in protein sequences. The web implementation of the algorithm contains several new features and enhancements: (i) one universal scoring matrix which allows identification of AGO-binding proteins in sequences representing all organisms; (ii) reduction of false positive predictions by improved selectivity of the algorithm; (iii) graphical interface to easily browse the prediction results; and (iv) the option to submit a DNA sequence which will be automatically translated in six frames before running the prediction algorithm.

Availability: Freely available at: http://bioinfo.amu.edu.pl/agos/.

Contact: wmk@amu.edu.pl

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

Received on December 28, 2010; revised on February 8, 2011; accepted on March 3, 2011

1 INTRODUCTION

The WG/GW domains recruit Argonaute (AGO) proteins to distinct components of the eukaryotic RNA silencing pathways and are strictly required for gene silencing by RNA interference (RNAi). The sequences of the domain are extremely divergent and are composed of quasi-repeated regions containing conserved tryptophan–glycine and/or glycine–tryptophan motifs (henceforth called WG/GW motifs), which are essential for binding multiple molecules of AGO proteins (El-Shami *et al.*, 2007). The hypervariable sequences of WG/GW domain are generally not alignable as positional homology cannot be precisely determined. Consequently, all commonly available bioinformatic tools (e.g. PSI-BLAST, HMMER and Gibbs sampler) fail to identify the AGO-binding domains in systematic analyses. There is presently no publicly available software specifically designed for detection of these sequence-divergent, but functionally conserved AGO-binding domains.

In a previous study (Karlowski *et al.*, 2010), we demonstrated by an *in silico* domain-swapping simulation between plant and mammalian WG/GW proteins that the amino acid composition of the AGO-binding sites is conserved. We designed a computational method for WG/GW domain detection based on the preferences of certain amino acids found within the plant WG/GW domains. In our screen we were able to identify all of the already characterized AGO-interacting proteins in plants (KTF1/SPT5, SPT5-like and NRPE1) and mammals (GW182 family members) as well as several other candidate WG/GW domain-containing proteins. The experimental verification of one of the proteins (a putative oxidoreductase), confirmed its AGO-binding capabilities. However, several proteins identified during the screening of animal genomes represented false predictions. These fragments represent low-complexity sequences rich in amino acid located at the top of our scoring table. The presence of such proteins which most likely have no AGO-binding activity was probably the result of a limited capability of the original plant scoring matrix to distinguish between the genuine WG/GW proteins and other compositionally biased molecules.

We present Agos (Argonaute-binding domain screener), a new web application for *de novo* identification of WG/GW Argonaute-binding domains in eukaryotic proteins. To overcome the restrictions of the original method, during the development of Agos we introduced several method improvements: (i) the initial sequence dataset of both plant and animal, experimentally confirmed, AGO-binding proteins; (ii) a new scoring table for all 400 possible combinations of dipeptides (this step significantly improved specificity of WG/GW domain identification and enhanced precise the annotation of domain boundaries); and (iii) a scoring matrix which now reflects compositional differences between the domain and the whole corresponding proteome. In this way the compositional signal of the domain is specifically amplified, leading to precise detection of conserved features of the domain.

2 METHODS

The detailed methodology of AGO-binding domain identification was described recently (Karlowski *et al.*, 2010). The initial sequence dataset was extended to cover a manually selected collection of plant and animal

*To whom correspondence should be addressed.

WG/GW protein domains and their orthologs with experimentally confirmed biological function. The sequence collection included NRPE1 (El-Shami *et al*., 2007), SPT5/KTF1 (He *et al*., 2009), SPT5-like (Bies-Etheve *et al*., 2009), WGRP1 from *Arabidopsis* (Karlowski *et al*., 2010), WAG1, cnjB from *Tetrahymena thermophila* (Bednenko *et al*., 2009), and GW182 from fly and GW182 family-related proteins from human and other mammalians (Eulalio *et al*., 2009).

The calculation of scoring tables (*dos* and *ics*; see Supplementary Material for details) and selection corresponding threshold values were already explained in the previous study. *Dos* scoring matrix was calculated based on the frequency distribution of all 400 combinations of dipeptides present in the initial dataset of WG/GW domains and corresponding proteomes. The *ics* scoring table represents the properties of the group of domains selected as the initial dataset for this analysis. In contrast to the *dos* score, where higher values represent better candidates, values tending towards zero represent a closer compositional relationship with the reference dataset.

3 RESULTS

3.1 Design and implementation

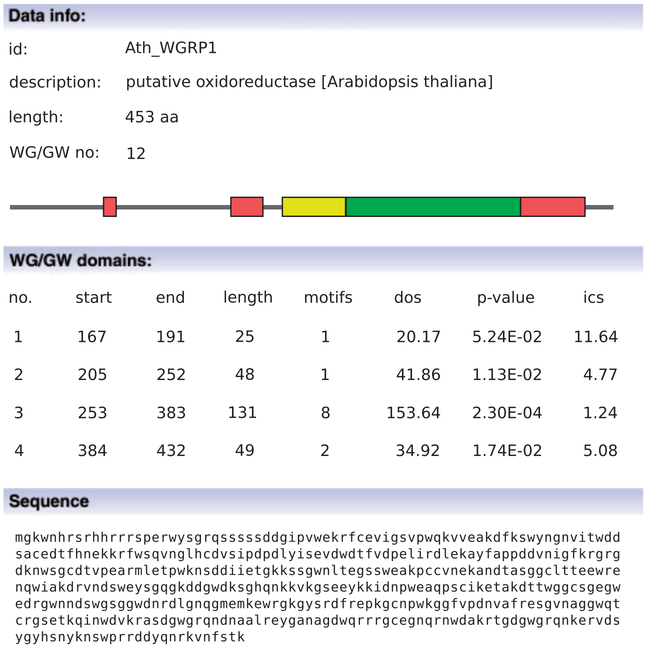
The computational engine of Agos was developed in Python and the web interface is implemented in Django. The tool is designed to accept one single protein or DNA sequence in FASTA format at a time. The DNA sequence is translated in all six reading frames before applying the domain detection algorithm. The query protein sequence must contain at least one WG or GW motif to be further processed. The data are posted to a WG/GW protein identification pipeline to screen for all regions containing WG/GW domains. The availability of two scoring systems, *dos* and *ics* (see Section 2) provides a measurement of the degree of compositional compatibility of the new domains with the already-confirmed AGO-binding proteins. Program output can either be displayed in the web browser window (interactive mode) or saved to a file (text/plain output). The interactive output mode uses standard DHTML (HTML, CSS and JavaScript), and no additional plugins are required.

3.2 Output of WG/GW domain identification results

The output report is divided into three separate fields: ‘Data info’, ‘WG/GW domains’ and ‘Sequence’ (Fig. 1).

The ‘Data info’ field provides basic information about the query sequence (id, description, length and number of WG/GW motifs), as well as a graphical view of the protein with marked positions for all detected potential WG/GW regions. The quality of GW domain predictions are color coded. Green blocks indicate domains that passed statistical threshold values (*dos* and *ics*). Yellow blocks label sequence regions that passed only *dos* score threshold. Red blocks correspond to regions having very low-compositional compatibility to the GW AGO-binding domain.

The ‘WG/GW domains’ field provides more detailed, textual information in the form of a table listing all the predicted domains sorted according to the start position in the protein. For each domain, the index number, the start and stop positions in the query sequence, the length of the domain, the number of WG/GW motifs, *dos* score, *P*-value and *ics* score are shown.



Data info:

id: Ath_WGRP1

description: putative oxidoreductase [Arabidopsis thaliana]

length: 453 aa

WG/GW no: 12

WG/GW domains:

no.	start	end	length	motifs	dos	p-value	ics
1	167	191	25	1	20.17	5.24E-02	11.64
2	205	252	48	1	41.86	1.13E-02	4.77
3	253	383	131	8	153.64	2.30E-04	1.24
4	384	432	49	2	34.92	1.74E-02	5.08

Sequence

```
mgkwnhrsrhhrrrsperwysgrqsssssddgipvwekrfcevigsvpwqkvveakdfkswyngnvitwdd
sacedtfhnekkrfwsqvnglhcdvsipdpdlyisevdwdtfvdpelirdlekayfappddvnigfkrgrg
dknwsgcdtvpearmletpwknsddiietgkkssgwnltegssweakpccvnekandtasggcltteewre
nqwiakdrvndsweysgqgkddgwdksghqnkkvkgseeykkidnpweaqpsciketakdttwggcsgegw
edrgwnndswgsggwdnrdlgnqgmemkewrgkgysrdfrepkgcnpwkggfvpdnvafresgvnaggwqt
crgsetkqinwdvkrasdgwgrqndnaalreyganagdwqrrrgcegnqrnwdakrtgdgwgrqnkervds
ygyhsnyknswprrddyqnrkvnfstk
```

Fig. 1. Results generated by Agos server.

The HTML results view provides additional interactive WG/GW domain match viewer and feature highlighting. Moving the cursor over a match block (with JavaScript option enabled) in the graphical protein view will highlight its position in the full-length protein sequence as well as the corresponding row in the ‘WG/GW domains’ table. Highlighted regions are preserved as long as the user does not move the cursor over another block. The plain text output provides no web links, but is optimal for copy/pasting and corresponds to the ‘WG/GW domains’ field where values are tab separated.

The ‘Sequence’ section displays the full-length query sequence.

Funding: Faculty of Biology Adam Mickiewicz University.

Conflict of Interest: none declared.

REFERENCES

Bednenko,J. *et al*. (2009) Two GW repeat proteins interact with Tetrahymena thermophila argonaute and promote genome rearrangement. *Mol. Cell. Biol.*, **29**, 5020–5030.

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.

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

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

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