MAGA: A Supervised Method to Detect Motifs From Annotated Groups in Alignments

Evolutionary Bioinformatics
Volume 16: 1–4
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934320916199

Pablo Mier and Miguel A Andrade-Navarro

Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, Mainz 55128, Germany

ABSTRACT: Multiple sequence alignments are usually phylogenetically driven. They are studied in the framework of evolution. But sometimes, it is interesting to study residue conservation at positions unconstrained by evolutionary rules. We present a supervised method to access a layer of information difficult to appreciate visually when many protein sequences are aligned. This new tool (MAGA; http://cbdm-01.zdv.uni-mainz.de/~munoz/maga/) locates positions in multiple sequence alignments differentially conserved in manually defined groups of sequences.

KEYWORDS: Sequence analysis, multiple sequence alignments, computational biology, web services, motif finding

RECEIVED: January 10, 2020. **ACCEPTED:** March 10, 2020.

TYPE: Methods and Protocols

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Deutsche Forschungsgemeinschaft (AN735/4-1 to M.A.A.-N.).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Pablo Mier, Institute of Organismic and Molecular Evolution, Johannes Gutenberg University Mainz, Hanns-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany. Email: munoz@uni-mainz.de

Introduction

Protein sequence alignments are based on the comparison of residues to detect similarities, which usually leads to conclusions about whether residues are conserved, mutated, deleted, or inserted in evolution. Tools like Clustal Omega,[1] T-Coffee,[2] MAFFT,[3] and MUSCLE[4] are some of the algorithms used daily by researchers to align sequences into multiple sequence alignments (MSAs). Here, we present a tool that postprocesses such alignments to facilitate the inference of residue properties specific to sequence groups defined by the user. This is a crucial step in the use of protein sequence alignments for the prediction of protein function.

The idea of locating specificity-determining sites (SDSs) by subfamily analysis of the sequences dates back to 1993.[5] Many other approaches have been developed since then following this line of thought.[6-8] All of them use unsupervised approaches that automatically place the sequences in groups by analyzing the SDSs, some also taking into consideration local physicochemical properties and phylogenetic information.

For exploratory purposes, we believe that a user-supervised approach, where one should be able to adapt the groups considered, can be of great help. Consider, for example, a situation in which the user has functional information of the proteins of a family that is independent of the phylogeny. Examining the conservation in groups of proteins defined by the user may reveal residues associated with that specific functional information.

In this article, we describe the MAGA (Motifs from Annotated Groups in Alignments) web tool, a simple way to infer conservation information from manually defined groups of sequences in alignments. We also provide 2 case studies to exemplify its use and showcase how it can provide meaningful results to researchers in a transversal way, not necessarily phylogenetically based.

Main Text

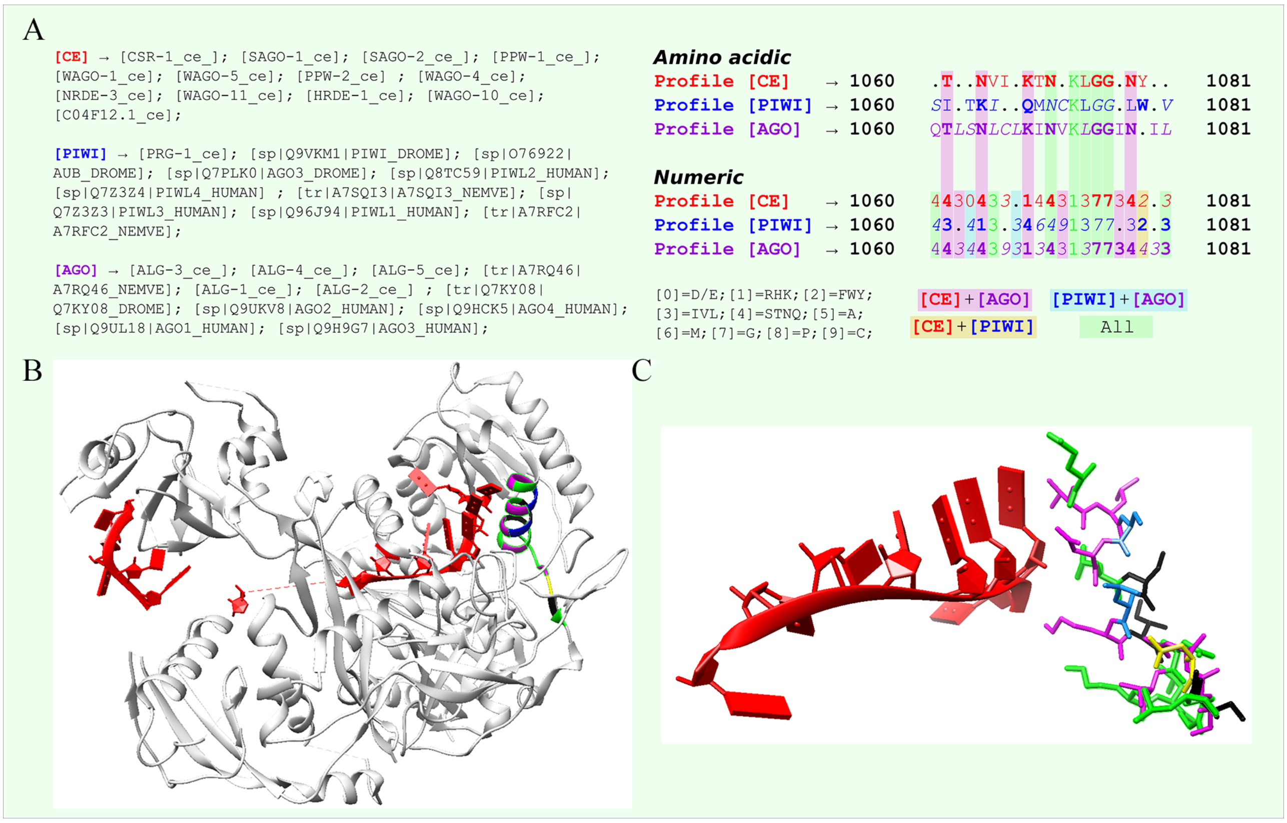
Workflow

MAGA takes as input an MSA of protein sequences, in FASTA format. In a first iteration, MAGA preprocesses the alignment and generates a profile with the shared residues from all input sequences, as a normal aligner would do. Then, it lets the user allocate the sequences in up to 6 groups and label them. Group assignments can be done manually or by uploading a file with assignments. The file is required to have 1 row per assignment with 2 columns indicating an ID and its assigned group identified by a number, respectively. Next, MAGA produces a profile per group with both the shared residues and group-conserved residues colored depending on the assigned group. To consider a residue at a position as conserved in a group, it must meet the following conditions:

1. There must be an amino acid that is more prevalent than all the other amino acids together for that group in that position.
2. The most prevalent amino acid must be more prevalent than the gaps in that group at that position.

The color code of the profiles displaying the group-conserved amino acids follows the notation:

- Conserved in all sequences → Colored in green.
- Per group, conserved in >50% but <75% of the sequences → Colored in red, blue, indigo, orange, violet or black, depending on the group.
- Per group, conserved in ⩾75% but <100% of the sequences → In italics and colored in red, blue, indigo, orange, violet, or black, depending on the group.



```
[CE] → [CSR-1_ce_]; [SAGO-1_ce]; [SAGO-2_ce_]; [PPW-1_ce_];
[WAGO-1_ce]; [WAGO-5_ce]; [PPW-2_ce] ; [WAGO-4_ce];
[NRDE-3_ce]; [WAGO-11_ce]; [HRDE-1_ce]; [WAGO-10_ce];
[C04F12.1_ce];

[PIWI] → [PRG-1_ce]; [sp|Q9VKM1|PIWI_DROME]; [sp|O76922|
AUB_DROME]; [sp|Q7PLK0|AGO3_DROME]; [sp|Q8TC59|PIWL2_HUMAN];
[sp|Q7Z3Z4|PIWL4_HUMAN] ; [tr|A7SQI3|A7SQI3_NEMVE]; [sp|
Q7Z3Z3|PIWL3_HUMAN]; [sp|Q96J94|PIWL1_HUMAN]; [tr|A7RFC2|
A7RFC2_NEMVE];

[AGO] → [ALG-3_ce_]; [ALG-4_ce_]; [ALG-5_ce]; [tr|A7RQ46|
A7RQ46_NEMVE]; [ALG-1_ce_]; [ALG-2_ce_] ; [tr|Q7KY08|
Q7KY08_DROME]; [sp|Q9UKV8|AGO2_HUMAN]; [sp|Q9HCK5|AGO4_HUMAN];
[sp|Q9UL18|AGO1_HUMAN]; [sp|Q9H9G7|AGO3_HUMAN];
```

```
Amino acidic
Profile [CE]   → 1060   .T..NVI.KTN.KLGG.NY..    1081
Profile [PIWI] → 1060   SI.TKI..QMNCKLGG.LW.V    1081
Profile [AGO]  → 1060   QTLSNLCLKINVKLGGIN.IL    1081

Numeric
Profile [CE]   → 1060   4430433.144313773 42.3   1081
Profile [PIWI] → 1060   43.413.346491377.32.3    1081
Profile [AGO]  → 1060   443443931343137734433    1081

[0]=D/E;[1]=RHK;[2]=FWY;     [CE]+[AGO]   [PIWI]+[AGO]
[3]=IVL;[4]=STNQ;[5]=A;      [CE]+[PIWI]     All
[6]=M;[7]=G;[8]=P;[9]=C;
```

Figure 1. Case study 1: Argonaute family. Results obtained in MAGA when using as input an MSA with 34 sequences from 3 related protein subfamilies (CE, PIWI, and AGO) and executed with amino acids and with categories. (A) Sequences are grouped in MAGA based on their subfamily, and conserved residues in the alignment positions 1060-1081 are highlighted depending on the groups in which they are conserved: [CE] + [AGO] (pink), [PIWI] + [AGO] (blue), [CE] + [PIWI] (yellow), and all (green). (B) Structure of human protein AGO3 (PDB:5VM9); in red, RNA chain; in colors based on their conservation, alignment positions 1060-1081. (C) Detailed interaction RNA-helix, with [CE] + [AGO] conserved residues (pink) pointing toward the RNA chain. MAGA indicates Motifs from Annotated Groups in Alignments; MSA, multiple sequence alignment; RNA, ribonucleic acid.

- Per group, conserved in all the sequences → In bold and colored in red, blue, indigo, orange, violet, or black, depending on the group.

The group assignment of the sequences can be iteratively modified. As an example, in a first iteration, one may cluster the sequences [A, B, C, D, E, F] based on taxonomy as [A, B, C] + [D, E, F], then differently based on a phenotype as [A, C] + [B, E] + [D, F], and finally, based on whether they present a sequence feature or not as [A, E, F] + [B, C, D]. Previous results are shown directly below the current results so that the comparison between the results originated from different arrangements can be easily done.

MAGA offers the possibility to convert the amino acids in the alignment to categories, to detect regions with similar physicochemical properties. The considered equivalences are as follows: [0] = D/E (negatively charged); [1] = RHK (positively charged); [2] = FWY (aromatic); [3] = IVL (aliphatic); [4] = STNQ (with polar uncharged side chains); [5] = A; [6] = M; [7] = G; [8] = P; and [9] = C. One can even use the same grouping with and without this feature, to compare their outputs, as shown in the section "Case study 1: Argonaute protein family."

Case study 1: Argonaute protein family

We prepared an MSA of 34 sequences from the Argonaute protein family, from 3 subfamilies: AGO, PIWI, and CE (Supplemental File 1). The CE is an Argonaute *Caenorhabditis elegans*–specific subfamily, equally distant from the AGO and PIWI subfamilies.[9] We executed MAGA with the MSA and grouped the sequences based on their subfamily. We explored the results obtained with amino acids and with categories. The results of residue conservation in the alignment positions 1060-1081, within the Piwi domain, are particularly interesting (Figure 1). Although there are several positions conserved in all groups (in green), most importantly is the high conservation between the CE and AGO subfamilies (in pink) (Figure 1A). This region happens to fold in an alpha-helix structure that binds RNA (Figure 1B). A closer look at the helix shows that CE + AGO conserved residues are indeed the ones in contact with the RNA (Figure 1C). The results indicate a similar binding specificity for CE and AGO proteins, in this region, whereas PIWI proteins would interact differently. The yeast *Schizosaccharomyces pombe* has only 1 Argonaute protein[10] (UniProt: O74957), with similar residues to those of CE + AGO (data not shown). The CE subfamily may have

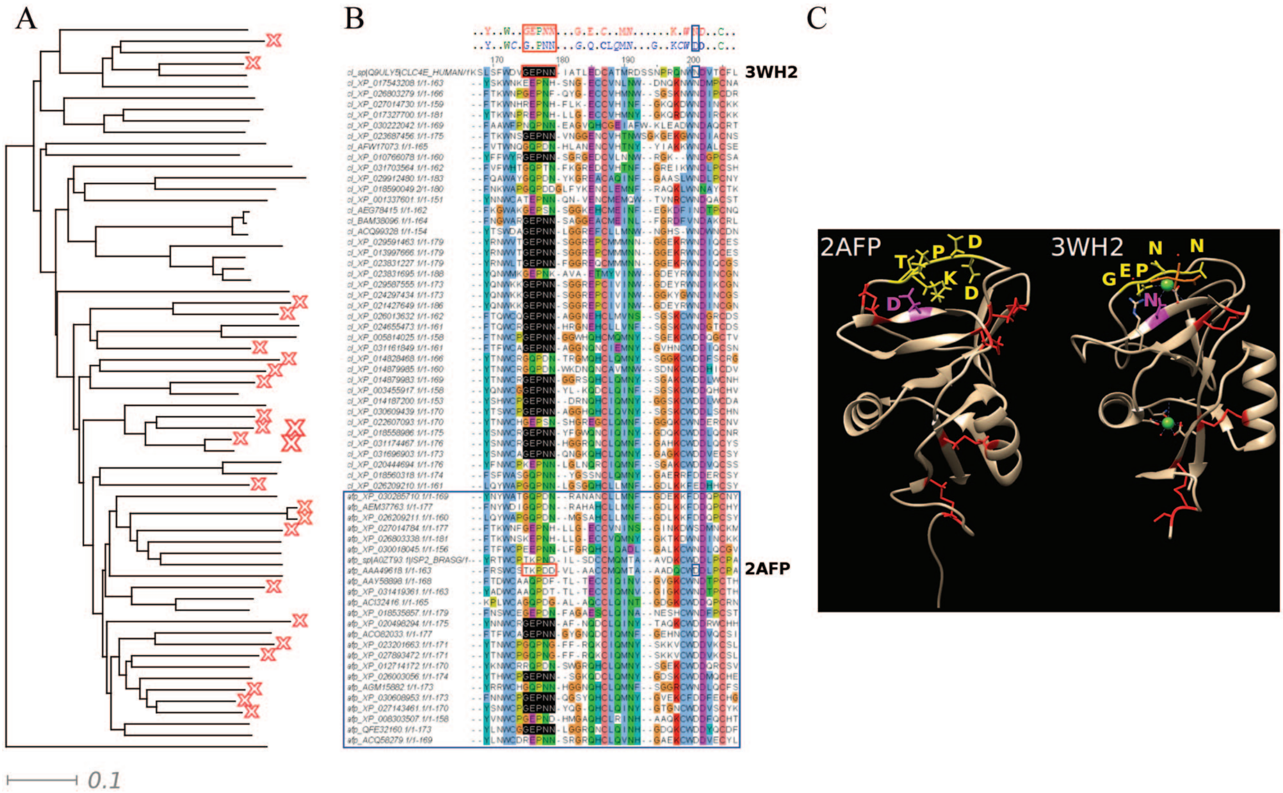


Figure 2. Case study 2: Evolution of fish anti-freeze proteins. Results obtained in MAGA when using as input an MSA with 64 proteins, which include 24 fish anti-freeze proteins (AFPs) and 40C-type 4 lectins (CLs). These AFPs evolved from duplicated CLs. (A) Phylogenetic tree derived from the MSA. AFPs are marked with red Xs. Their position in different branches indicates that they emerged in multiple independent events. The outlier at the bottom is the human C-type lectin domain family 4 member E. (B) Part of the MSA with CLs at the top and AFP at the bottom. Sequence identifiers have been added the label "CL" or "AFP" as a prefix for clarity. Residues are colored according to type and conservation and the GEPNN motif has been marked (using Jalview).[16] MAGA conservation computed for the region is shown above. Red boxes indicate the differentially conserved GEPNN motif, present in human C-type lectin domain family 4 member E (PDB:3WH2), which is less conserved in the AFPs (eg, it is "TKPDD" in the AFP from *H americanus*). Blue boxes indicate a differentially conserved residue, mostly N in the CLs and D in the AFPs. (C) Structural information for the *H americanus* AFP (PDB:2AFP) (left) and the human C-type lectin domain family 4 member E (PDB:3WH2) (right). Both proteins have a similar distribution of secondary structural motifs and conserved cysteine bridges (all cysteines are marked in red). But the human protein (right) has a pocket that binds citrate (in orange) using a Ca^{2+} atom (green ball, top) coordinated by the GEPNN sequence (in yellow) and an Asn (N, in magenta) in an opposite beta strand. These motifs are not conserved in AFPs (TKPDD in the AFP, left) or conserved differently (D, in magenta in the AFP, left), reflecting the absence of selection pressure to keep the sugar-binding pocket. MAGA indicates Motifs from Annotated Groups in Alignments; MSAs, multiple sequence alignments; RNA, ribonucleic acid.

initially appeared in *C elegans* as a duplication from an AGO protein, and the PIWI subfamily would have diverged then in how it interacts with the RNA in the studied region.

Case study 2: evolution of fish anti-freeze proteins

Fish have several families of anti-freeze proteins (AFPs). The type II family constitutes a suitable case study for MAGA because its members originated from multiple independent events of duplication and evolution from C-type 4 lectins.[11] Lectins are vertebrate proteins that bind carbohydrates, and in particular, C-type lectins have a carbohydrate recognition domain with 2 coordinated Ca^{2+} ions and 4 conserved cysteine bridges.[12] The carbohydrate binding is calcium-dependent because the binding pocket is constructed with one of the coordinated Ca^{2+} ions.[13] However, we would expect that the AFPs evolved from lectins will no longer need to bind carbohydrates and will lose the selective pressure to keep the pocket that recognizes the carbohydrate. To test this hypothesis with MAGA, we searched for fish homologs of the type II anti-freeze protein of the sea raven (*Hemitripterus americanus*), whose structure is known,[14] using the NCBI BLAST tool.[15] We included the human C-type lectin domain family 4 member E (CLEC4E), whose structure is also known,[13] and acted as an outgroup in the phylogenetic analysis. The resulting set of 64 sequences (Supplemental File 2) was used to construct an MSA. We categorized the sequences as CLs or AFPs according to their annotations. The phylogenetic tree (Figure 2A) displays AFPs scattered across multiple branches as a result of the multiple independent events of the evolution of duplicated CLs into AFPs. Using the MSA and the categorization into 2 groups in MAGA resulted in the identification of a motif "GEPNN" that was significantly more conserved in CLs than in AFPs (Figure 2B, red box). This motif was present in 22 out of 40 CLs (55%) and in 6 out of 24 AFPs (24%). It was noted that mutations of the "EPN" sequence in human CLEC4E might disrupt the Ca^{2+} ion-mediated binding network that recognizes the carbohydrate.[13] Another differential feature is a position that displays predominantly Asn or Asp in CLs and AFPs, respectively (Figure 2B, black box). Comparison of the structures of sea raven type II AFP and human CLEC4E indicates that the general structure including cysteine bridges is conserved (Figure 2C), but the GEPNN motif (yellow) or the opposed Asn (magenta) present in human is different in the fish protein, which fills the room corresponding to the Ca^{2+} ion in the human protein with amino acid side chains. In this case, even relying on sequence annotations that might be inaccurate if based on homology and not on experiments, MAGA can be used to identify sequence features common to AFPs that evolved independently within the C-type 4 lectin family.

Conclusions

Amino acid conservation in an MSA reflects the phylogenetic proximity of the sequences, functionality, or both. The idea behind MAGA (http://cbdm-01.zdv.uni-mainz.de/~munoz/maga/) is to find positions conserved due to functionality that are not necessarily phylogenetically related. Taxonomic information in this respect is still important to rule out the evolutive component when detecting a conserved residue in a manually-defined group of sequences.

Although it is trivial to detect group-conserved residues or regions when working with small MSA, the task becomes much harder for larger MSA and when analyzing conserved physicochemical properties in a position, not just amino acids, which are more difficult to detect visually. The MAGA web tool addresses this problem allowing the detection of sequence conservation in a transversal nonphylogenetically driven way. We strongly believe it will help users to detect meaningful biological functions and motifs in an exploratory analysis.

Acknowledgements

The authors thank Miguel Almeida and René Ketting (IMB/JGU, Mainz) for insightful discussions regarding the Argonaute protein family.

Author Contributions

P.M. and M.A.A.-N. conceived the project. P.M. developed and implemented MAGA. M.A.A.-N. generated and analyzed the case studies. M.A.A.-N. supervised the project. P.M. and M.A.A.-N. drafted the manuscript, and read and approved the final manuscript.

Availability of data and materials

The MAGA web tool is available at http://cbdm-01.zdv.uni-mainz.de/~munoz/maga/, with no restrictions for users. The multiple sequence alignments used in the case studies are available as Supplementary Files.

ORCID iDs

Pablo Mier https://orcid.org/0000-0003-3663-2352
Miguel A Andrade-Navarro https://orcid.org/0000-0001-6650-1711

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
2. Di Tommaso P, Moretti S, Xenarios I, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*. 2011;39:W13-W17.
3. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2019;20:1160-1166. doi:10.1093/bib/bbx108.
4. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797.
5. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci*. 1993;9:745-756.
6. Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res*. 2004;32:W424-W428.
7. Ye K, Feenstra A, Heringa J, Ijzerman AP, Marchiori E. Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. *Bioinformatics*. 2008;24:18-25.
8. Chakraborty A, Chakrabarti S. A survey on prediction of specificity-determining sites in proteins. *Brief Bioinform*. 2015;16:71-88.
9. Yigit E, Batista PJ, Bei Y, Pang KM, Chen CC, Tolia NH. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell*. 2006;127:747-757.
10. Smialowska A, Djupedal I, Wang J, Kylsten P, Swoboda P, Ekwall K. RNAi mediates post-transcriptional repression of gene expression in fission yeast *Schizosaccharomyces pombe*. *Biochem Biophys Res Commun*. 2014;444:254-259.
11. Fletcher GL, Hew CL, Davies PL. Antifreeze proteins of teleost fishes. *Annu Rev Physiol*. 2001;63:359-390.
12. Zelensky AN, Gready JE. The C-type lectin-like domain superfamily. *FEBS J*. 2005;272:6179-6217.
13. Furukawa A, Kamishikiryo J, Mori D, et al. Structural analysis for glycolipid recognition by the C-type lectins Mincle and MCL. *Proc Natl Acad Sci USA*. 2013;110:17438-17443.
14. Gronwald W, Loewen MC, Lix B, et al. The solution structure of type II antifreeze protein reveals a new member of the lectin family. *Biochemistry*. 1998;37:4712-4721.
15. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2016;44:D7-D19.
16. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25:1189-1191.