BMC Structural Biology

Software

GANDivAWeb: A web server for detecting early folding units ("foldons") from protein 3D structures

Thomas Laborde[1], Masaru Tomita[2] and Arun Krishnan*[2]

Address: [1]ENSIERB, Bordeaux, France and [2]Institute for Advanced Biosciences, Keio University, 14-1, Baba-Cho, Tsuruoka, Yamagata-ken, 997-0035, Japan

Email: Thomas Laborde - laborde@enseirb.fr; Masaru Tomita - mt@sfc.iab.keio.ac.jp; Arun Krishnan* - drarunkrishnan@gmail.com

* Corresponding author

Published: 7 March 2008

BMC Structural Biology 2008, **8**:15 doi:10.1186/1472-6807-8-15

Received: 18 October 2007
Accepted: 7 March 2008

This article is available from: http://www.biomedcentral.com/1472-6807/8/15

Abstract

Background: It has long been known that small regions of proteins tend to fold independently and are then stabilized by interactions between these distinct subunits or modules. Such units, also known as autonomous folding units (AFUs) or"foldons" play a key role in protein folding. A knowledge of such early folding units has diverse applications in protein engineering as well as in developing an understanding of the protein folding process. Such AFUs can also be used as model systems in order to study the structural organization of proteins.

Results: In an earlier work, we had utilized a global network partitioning algorithm to identify modules in proteins. We had shown that these modules correlate well with AFUs. In this work, we have developed a webserver, GANDivAWeb, to identify early folding units or "foldons" in networks using the algorithm described earlier. The website has three functionalities: (a) It is able to display information on the modularity of a database of 1420 proteins used in the original work, (b) It can take as input an uploaded PDB file, identify the modules using the GANDivA algorithm and email the results back to the user and (c) It can take as input an uploaded PDB file and a results file (obtained from functionality (b)) and display the results using the embedded viewer. The results include the module decomposition of the protein, plots of cartoon representations of the protein colored by module identity and connectivity as well as contour plots of the hydrophobicity and relative accessible surface area (RASA) distributions.

Conclusion: We believe that the GANDivAWeb server, will be a useful tool for scientists interested in the phenomena of protein folding as well as in protein engineering. Our tool not only provides a knowledge of the AFUs through a natural graph partitioning approach but is also able to identify residues that are critical during folding. It is our intention to use this tool to study the topological determinants of protein folding by analyzing the topological changes in proteins over the unfolding/folding pathways.

Background

It is recognized that small regions of proteins tend to fold independently and are then stabilized by interactions between these distinct subunits or modules. The dissection of proteins into structurally independent and functionally distinct subunits led to the idea that proteins can be considered as collections of smaller units such as domains [1]. Different definitions of domains have been

in existence [2]. While some define a domain as a recognizable substructure within a protein and connected to other domains by very few structural elements such as a loop or a helix, others define domains as a parts of a protein molecule that behave in a quasi-independent manner and are considered as cooperative units in protein folding [3,4]. A further definition describes a domain as a relatively compact part of a protein that is characterized by its own pattern of intramolecular collective dynamics and which are distinguishable from those of other domains [5-8]. In the belief that an unequivocal definition of a module must be based on the most fundamental property of protein 3D structure, namely, the adjacency matrix of inter-residues contact, we adopted a network representation of the protein.

In an earlier work [9], we had used a well-established, global method for identifying modules in networks [10]. The algorithm converges towards the maximization of the modularity of the given protein network; the network being defined by an adjacency matrix with a 1 denoting the existence of residue-residue contacts and a 0 for non-contacts. Maximizing the modularity score, as defined by Guimera *et al*. [10] results in maximizing intra-module contacts while minimizing the inter-module contacts. The modularity *M* is given by

$$M = \sum_{s=1}^{N_m} \frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2 \quad (1)$$

where N_m is the number of modules, L is the number of links in the network, l is the number of links between nodes in module s and d is the sum of the degrees of the node in module s. In doing so, this allows the representation of the residues of the protein in terms of their intra-module degree, z and participation coefficient, P., which are given by

$$z_i = \frac{\kappa_i - \bar{\kappa}_{s_j}}{\sigma_{\kappa_{s_j}}}$$
$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i}\right)^2 \quad (2)$$

where κ_i is the number of links of residue i to other residues in its module s_i, $\bar{\kappa}_{s_j}$ is the average of κ over all residues in module s_j, $w_{k_{s_j}}$ is the standard deviation of κ in module s_j, κ_{is} is the number of links of node i to nodes in module s and k_i is the total degree of node i.

We demonstrated that the labeling of residues in terms of these invariants, allowed for information rich representations of the studied proteins as well as to sketch a new way to link sequence, structure and the dynamical properties of proteins. We discovered a strong invariant character of protein molecules in terms of P/z characterization, pointing to a common topological design of all protein structures. This invariant representation, applied to different protein systems enabled us to identify the possible functional role of high P/z residues during the folding process. Effectively, this invariance is a cartographic representation of the contact network for proteins and is represented by the plot of the residues in the P - z space. Since it is identical for all the proteins, it does not embed any structural peculiarities or information for separating between different protein folds [11,12].

We also observed that the modules identified using the procedure outlined above correlated well with early folding units or "foldons" and thus a knowledge of the modules existing in a given protein can help to identify residues that are critical for folding.

A significant use for the modules identified using our methodology is for the development of algorithms for protein 3D structure determination. In addition knowing the modules for a protein can help in the understanding of the folding pathway for that protein since residues with high $|P/z|$ values tend to be protected during transition state and hence are fixed early in the folding process.

The modules can also be used for engineering new enzymes which is typically carried out by building a chimera of multiple proteins by cutting and pasting sequences from the respective proteins. A knowledge of the modules can guide the cuts in order to obtain chimeras that can fold in-vitro. In addition models of such early folding units can be invaluable in understanding the biochemical pathways of diseases that are known to be pathological through partially folded forms of proteins leading to the development of therapeutics.

Implementation and Results

GANDivAWeb

The GANDivAWeb webserver is based on the GANDivA (Genetic Algorithm-based Network modularity DetectIVe Algorithm) algorithm which is an implementation of the Guimera algorithm [10] that uses a genetic algorithm (GA) to optimize the modularity score, instead of simulated annealing which was used in the original algorithm. The algorithm has been written in C and was developed using the parallel genetic algorithm PGAPACK [13], the GNU Scientific Library [14] and the OpenMPI [15] and MPICH [16] MPI [17] implementations. The webserver has three main functionalities:

1. It displays the results from the application of the modularity algorithm GANDivA on a set of 1420 single-chain, globular proteins that were used as the basis for the work done in [9].

2. The main functionality of the website however, is the ability to process an uploaded protein structure file (in PDB format) using GANDivA and to send the results back to the user. Since GANDivA is a stochastic algorithm, the algorithm is run a number of times and the best results, as denoted by the largest modularity score obtained, is emailed back to the user. The user can set certain parameters like the maximum number of modules to be determined, the number of times to run the algorithm, the number of generations for the genetic algorithm as well as the number of generations for the fitness score to remain constant before it is assumed to have converged to a solution. An algorithm that takes into account the size of the protein (number of amino acid residues) as well as the number of jobs in the queue for the cluster, notifies the user about the expected time required for the completion of the job. The results emailed back to the user are made up of the following parts:

• A results file that contains the details of the modular decomposition of the protein. The results file 1(A) contains the modularity score in the first line. This is succeeded by the following columns: residue number, module number, intra-module connectivity (z), inter-module participation coefficient (P), P/z, $|P/z|$. The second column indicates the module to which each residue belongs. In addition to the results file, five different figures are generated.

• Along with the results file four figures are also included in the final results. The first two figures (Figures 1(B) and 1(C)) show cartoon representations of the protein with the residues colored according to the modules that they belong to and the $|P/z|$ value, respectively. High $|P/z|$ valued residues act as structural stabilizers and have been found to be correlated with the residues that are protected in the transition phase.

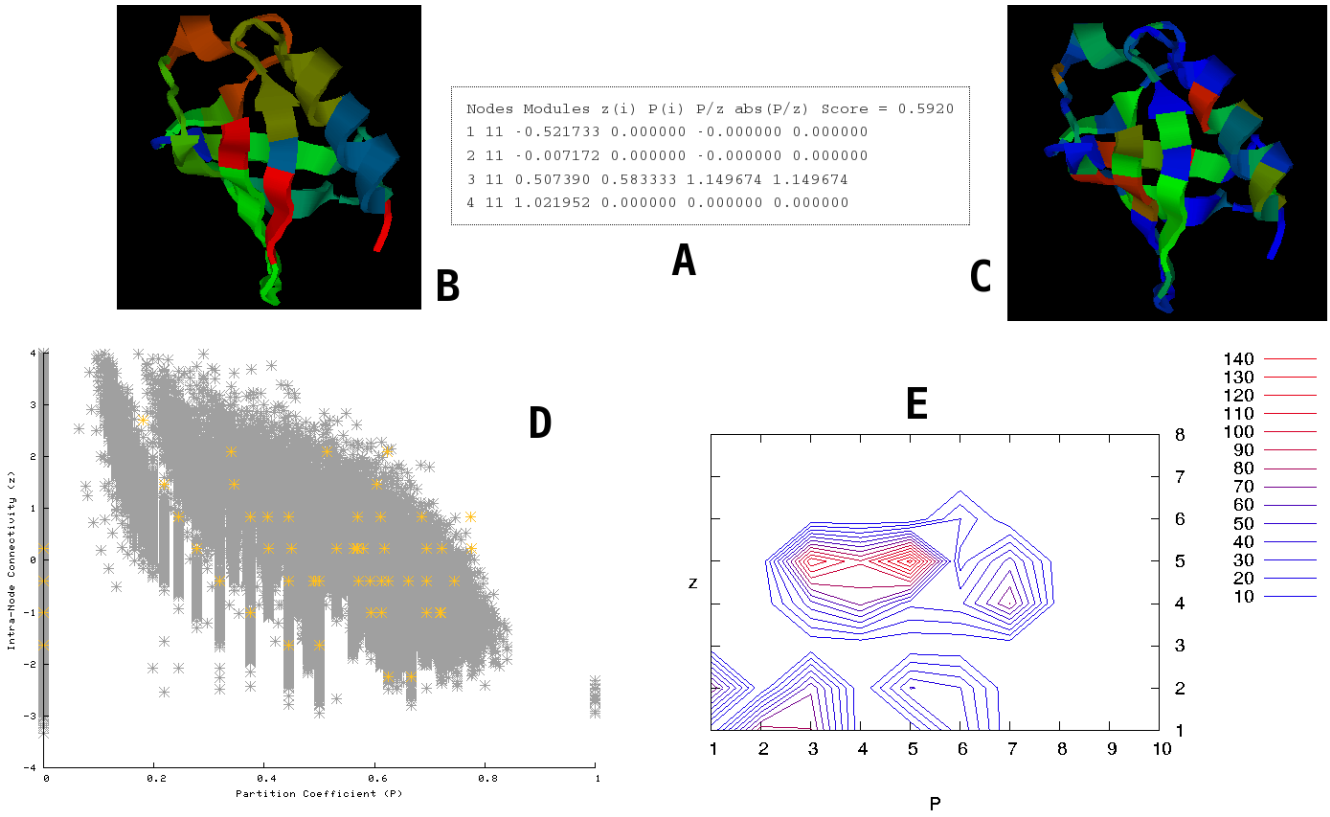

Figure 1

The figure shows the results emailed back to the user. (A): The results file, (B): Cartoon representation of the protein with the residues colored according to the modules, (C): Cartoon representation of the protein with the residues colored according to the $|P/z|$ value, (D): Scatter plot of the residues on the P - z plane, (E): Contour plots of the distribution of the buried and surface residues (Relative Accessible Surface Area (RASA) plot). See text for detailed explanations of the different figures.

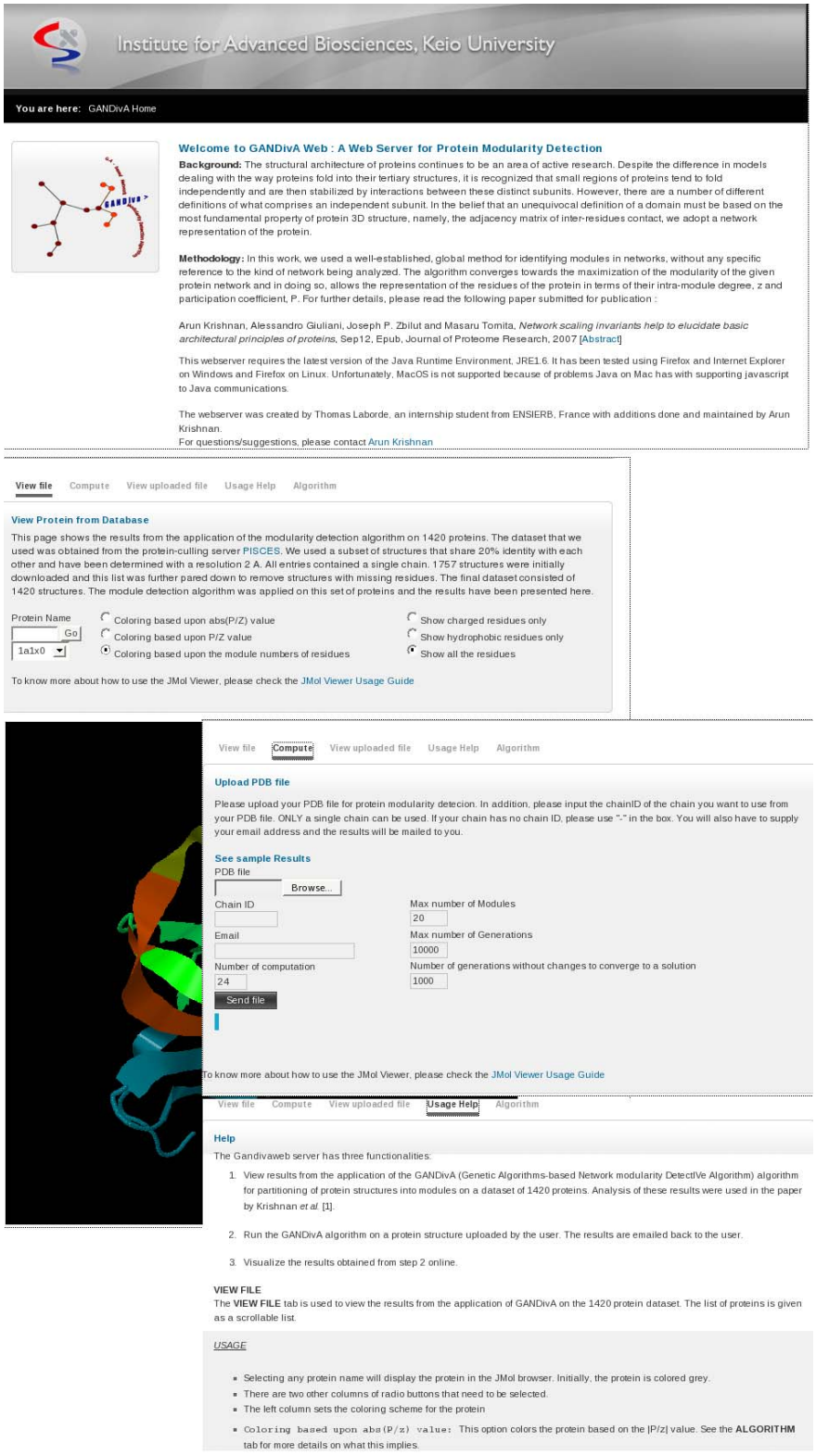

Figure 2
The GANDivAWeb User Interface: The figure shows the user interface for gandivaweb with snapshots of the different tabs corresponding to the main functionalities shown.

• The third figure (Figure 1(D) shows a scatter plot of the residues on the P - z plane. This scatter plot is overlaid on the scatter plot of the residues for the 1420 proteins studied in [9]. Marked deviation from the typical "dentist's chair" shape for a particular protein suggests non-native contacts.

• The fourth (Figure 1(E) and fifth (figure not shown) figures show contour plots of the distribution of the buried and surface residues (Relative Accessible Surface Area (RASA) plot) and the hydrophobic and polar residues over the topological P - z space, respectively. The distributions were obtained by first calculating the residue accessible surface area (RASA) and hydrophobicities respectively for each residue followed by partitioning the P - z space into an 8 × 10 grid and then calculating the mean value of the RASA and hydrophobicities in each bin. The mean value for each cell of the 8 × 10 grid was then plotted as a contour plot. Significant differences have been observed between these distributions for native and decoy proteins, with native distributions showing more structure (inclusive of two clearly defined regions where hydrophobic/low RASA residues are embedded).

3. The third functionality of the webserver is in displaying the results obtained using GANDivA. The user can upload the PDB file and the results file (obtained from functionality 2 above) and view the results in the embedded JMol viewer. For both functionalities 1 and 3, the user can choose to view the protein colored by the modules that the residues belong to or by the $|P/z|$ values. The user can also choose to view only hydrophobic, only polar or all the residues. Additionally, the viewer can choose the graphical representations of the protein that are a part of the embedded viewer.

Figure 2 shows screenshots of the different tabs pertaining to the various functionalities of the GANDivAWeb server.

Discussion

The GANDivAWeb server has been designed to partition any given PDB structure into modules. The modules thus obtained correlate well with autonomous folding units. Additionally, high $|P/z|$ valued residues are identified. These residues have been shown [9] to correlate well with residues that are protected early during the folding process. A knowledge of such units can help in understanding the folding process. It can also be used for engineering enzymes as mentioned earlier. Enzymes are typically engineered by cutting and pasting sequences from multiple proteins. Knowing the "natural" boundaries of protein folding modules can guide the "cuts" required in order to engineer proteins that can fold in-vitro. Moreover, the knowledge of early folding units can help in understanding the biophysical and biochemical causes of diseases that are caused by the misfolding of proteins.

Conclusion

We believe that the GANDivAWeb server will be of immense use to scientists interested in the phenomena of protein folding and those studying the architectural and structural organization of proteins.

Additionally, the site will also be useful for scientists interested in the engineering of novel enzymes by providing them with a modularized view of the protein.

Availability

The webserver can be found at http://gandivaweb.iab.keio.ac.jp

Authors' contributions

AK wrote the main algorithm and wrote the paper. TL designed and implemented the webserver and also helped in the writing of the paper. MT was in charge of the overall project.

Acknowledgements

The authors would like to thank Bharath Krishnan for his help in improving the stability fo the web server. This project is supported by the Yamagata prefecture and Tsuruoka city research grants.

References

1. Fischer KF, Marquesee S: **A rapid test for identification of autonomous folding units in proteins.** *J Mol Biol* 2000, **302:**701-712.
2. Yesylevskyy SO, Kharkyanen VN, Demchenko AP: **Dynamic protein domains: Identification, Interdependence and Stability.** *BioPhysical Journal* 2006, **91:**670-685.
3. Sato S, Kuhlman B, Wu WJ, Raleigh DP: **Folding of the multidomain ribosomal protein L9: the two domains fold independently with remarkably different rates.** *Biochemistry* 1999, **38:**5643-5650.
4. Jaenicke R: **Stability and folding of domain proteins.** *Prog Biophys Mol Biol* 1999, **71:**155-241.
5. Wriggers W, Schulten K: **Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates.** *Proteins* 1997, **29:**1-14.
6. Hayward S, Berendsen HJ: **Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme.** *Proteins* 1998, **30:**144-154.
7. Hinsen K: **Analysis of domain motions by approximate normal mode calculations.** *Proteins* 1998, **33:**417-429.
8. Hinsen K, Thomas A, field MJ: **Analysis of domain motions in large proteins.** *Proteins* 1999, **34:**369-382.
9. Krishnan A, Giuliani A, Zbilut JP, Tomita M: **Network scaling invariants help to elucidate basic topological principles of proteins.** *Journal of Proteome Research* 2007, **6(10):**3924-3934.
10. Guimera R, Amaral LAN: **Functional cartography of complex metabolic networks.** *Nature* 2005, **433:**895-900.
11. Zbilut JP, Chua GH, Krishnan A, Bossa C, Rother K, Webber CL, Giuliani A: **A Topologically Related Singularity Suggests a Maximum Preferred Size for Protein Domains.** *Proteins: Struct Funct Bioin* 2007, **66(3):**621-629.
12. Zbilut JP, Chua GH, Krishnan A, Bossa C, Colafranceschi M, Giuliani A: **Entropic criteria for protein folding derived from recurrences: Six residues patch as the basic protein word.** *FEBS Letters* 2006, **580(20):**4861-4864.

13. Levine D: **Users guide to the PGAPack parallel genetic algorithm library.** 1996 [ftp://info.mcs.anl.gov/pub/pgapack/~pgapack.tar.Z].
14. Pierce R: **The gnu scientific software library.** 1996 [http://www.gnu.org/software/gsl/].
15. Gabriel E, Fagg GE, Bosilca G, Angskun T, Dongarra JJ, Squyres JM, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain RH, Daniel DJ, Graham RL, Woodall TS: **Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation.** *11th European PVM/MPI Users' Group Meeting, Budapest, Hungary* 2004.
16. Gropp W, Lusk E, Doss N, Skjellum A: **A high-performance, portable implementation of the MPI message passing interface standard.** *Parallel Computing* 1996, **22(6):**789-828.
17. **MPI: Message Passing Interface** [http://www-unix.mcs.anl.gov/mpi/]