Protein Engineering, Design & Selection vol. 20 no. 10 pp. 521–523, 2007
Published online August 24, 2007 doi:10.1093/protein/gzm042

SHORT COMMUNICATION

The PASTA server for protein aggregation prediction

Antonio Trovato[1,2], Flavio Seno[1,2,3] and Silvio C.E. Tosatto[4,5]

[1]Department of Physics 'G. Galilei', University of Padova, [2]CNISM, Padova Unit, [3]INFN, Sezione di Padova and [4]Department of Biology and CRIBI Biotech Centre, University of Padova, Viale G. Colombo 3, 35131 Padova, Italy

[5]To whom correspondence should be addressed.
E-mail: silvio.tosatto@unipd.it

Many different proteins aggregate into amyloid fibrils characterized by cross-β structure. β-strands contributed by distinct protein molecules are generally found in a parallel in-register alignment. Here, we describe the web server for a novel algorithm, prediction of amyloid structure aggregation (PASTA), to predict the most aggregation-prone portions and the corresponding β-strand inter-molecular pairing for a given input sequence. PASTA was previously shown to yield results in excellent agreement with available experimental observations, when tested on both natively unfolded and structured proteins. The web server and downloadable source code are freely accessible from the URL: http://protein.cribi.unipd.it/pasta/.

Keywords: amyloid fibrils/cross-beta structure/parallel in-register arrangement/protein aggregation

Introduction

Several neurodegenerative disorders in humans are associated with the conversion of peptides and proteins from their soluble functional forms into well-defined fibrillar aggregates, generally described as amyloid fibrils (Chiti and Dobson, 2006). Many other proteins, not related to protein deposition diseases, have been found to form amyloid-like fibrils *in vitro* (Chiti *et al.*, 2003), suggesting that the ability to form the amyloid structure is encoded in the backbone interactions of a generic polypeptide chain. Notably, amyloids are produced by some organisms for functional properties without deleterious effects (Fowler *et al.*, 2006). The emergence of a limited menu of native-like conformations for a single chain and of β-aggregate structures for multiple chains was rationalized theoretically within a sequence independent coarse-grained framework (Hoang *et al.*, 2006).

In spite of the great variability in both sequences and soluble-state structures of precursor proteins, the resulting fibrils exhibit common properties (Sunde and Blake, 1997). Experimental studies identified the regions of the sequence forming and stabilizing the cross-β core of the fibrils, and clarified the nature of the intermolecular contacts. Parallel in register arrangements (PIRA) of β-strands in the fibril core occurs quite frequently (Ferguson *et al.*, 2006), but anti-parallel arrangements are also possible (Makin *et al.*, 2005). Mutational studies of the amyloid aggregation kinetics revealed simple correlations between physico-chemical properties and aggregation propensities, allowing the development of different methods which successfully predict aggregation-prone regions [for a recent review see (Caflisch, 2006)]. All approaches focus on predicting the β-aggregation propensity of a sequence stretch by itself. A new algorithm, prediction of amyloid structure aggregation (PASTA), was recently introduced by editing a pair-wise energy function for residues facing one another within a β-sheet (Trovato *et al.*, 2006). Two different propensity sets were extracted depending on the orientation (parallel or anti-parallel) of the neighboring strands, from a dataset of known native structures of globular proteins. PASTA associates energies to specific β-pairings of two sequence stretches of the same length, and further assumes that distinct protein molecules involved in fibril formation will adopt the minimum energy β-pairings in order to better stabilize the cross-β core. A novel feature of PASTA is the ability to predict the registry of the inter-molecular hydrogen bonds formed between amyloidogenic sequence stretches. In this way, the observed tendency of several proteins towards PIRA was rationalized on general grounds. PASTA, however, has also the intrinsic possibility to predict not in register alignment exactly since it considers all the possible matches of the replicas of the same sequence. The good performance of PASTA was tested on both natively unfolded (Trovato *et al.*, 2006) and structured proteins (Trovato *et al.*, 2007).

Server description

The PASTA server takes an amino acid sequence as input and predicts which portions of the sequence are more likely to stabilize the cross-β core of fibrillar aggregates. The input form is very simple, and requires an email address and (optional) title for the prediction job. The output can be divided in three parts: top pairing energies, aggregation profile and pairing matrix.

The top pairing energies are shown in the central part of the output page. Each line contains a predicted high scoring pairing, complete with localization (i.e. residue numbers) and orientation (parallel or anti-parallel). The number of pairings to be output is set in the input form, with a default of 10 pairings. The PASTA energy is indicative of the aggregation propensity. Benchmarking performed on the dataset of 179 peptides derived from the literature (Fernandez-Escamilla *et al.*, 2004) revealed close to 80% true positive predictions with a ~20% false positive rate at a PASTA energy threshold of −4.0 (Fig. 1).

The aggregation profile and pairing matrix are provided through links to PDF files. The aggregation profile shows the normalized per-residue probability $h(k)$ calculated from Eq. (5) in Trovato *et al.*, (2006). It serves to give a visual representation of which regions of the sequence are more likely to aggregate. The pairing matrix is based on the normalized two-dimensional probability $h_2(k,m)$ [see Eq. (6) in Trovato *et al.*, (2006)] based on the self-alignment of the sequence.

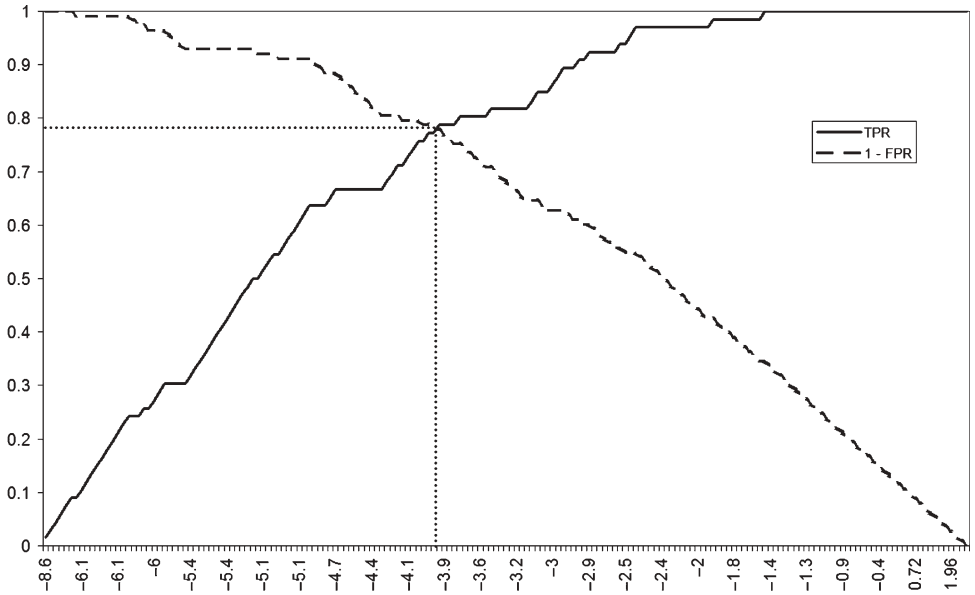

Fig. 1. Benchmarking of PASTA on a set of 179 experimentally verified peptides from the literature in terms of true positive rate (TPR) and false positive rate (FPR). Both TPR and (1—FPR) are plotted as a function of the PASTA energy threshold. The dotted line corresponding to a threshold of ~ -4.0 achieves the best compromise with $\sim$78% for both TPR and (1—FPR).

This can be seen as a β-pairing contact map highlighting, in increasing shades of grey, possible β-fibrillar topologies.

As an example, we show in Fig. 2 the PASTA output for the human amyloid β-peptide ($A\beta_{1-40}$), a peptide known to be involved in the Alzheimer's disease and other pathological conditions such as hereditary cerebral hemmorhage with amyloidosis and inclusion-body myositis (Chiti and Dobson, 2006). The two top-scoring pairings (residues 12–20 and 31–40, Fig. 2A) and the predicted PIRA alignment (Fig. 2C) are in very good agreement with experimental evidence (residues 12–24 and 30–40, varying somewhat between reports) (Petkova *et al.*, 2002).

Source code

For those wishing to run PASTA on large sequence ensembles (e.g. entire genomes) or interested in extending the approach, we are providing the source code as a downloadable TAR archive, reachable from the server homepage. The source code consists of an ANSI C program to compute the

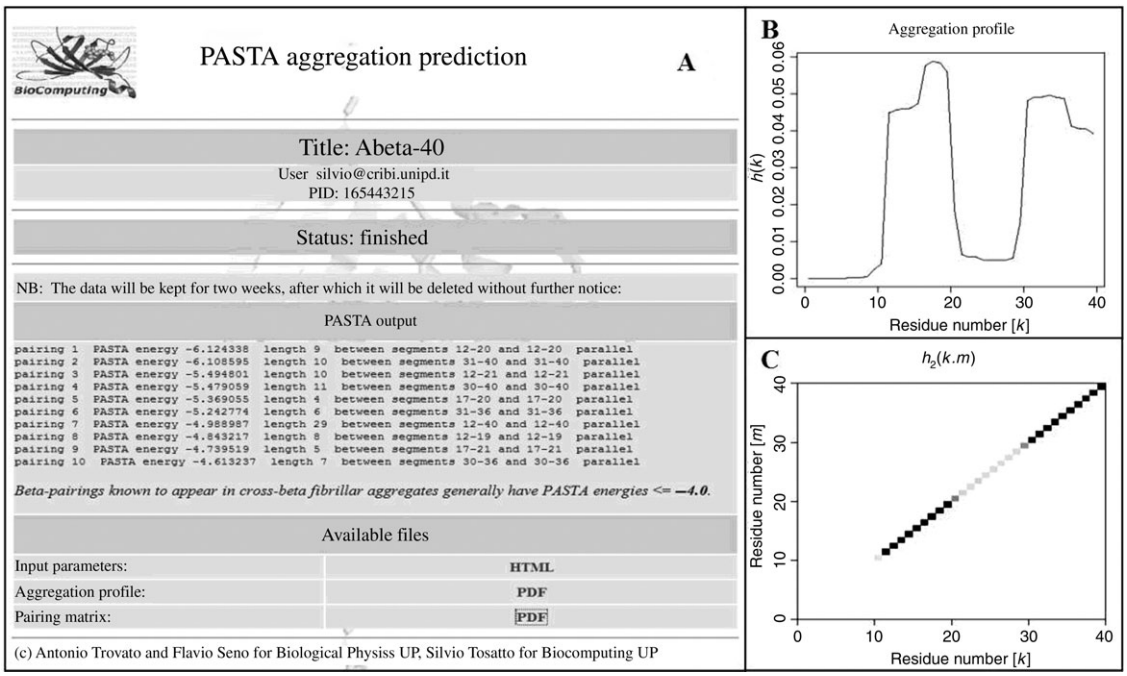

Fig. 2. Sample output of the PASTA server. (**A**) The main results window. This page contains the listing of the top 10 pairings, their energy and orientation (parallel or antiparallel) as well as links to the two remaining plots. (**B**) Aggregation plot, showing the aggregation probability as a function of amino acid position. (**C**) Pairing matrix, where the self-alignment of the sequence is used to indicate pairing probability in increasing shades of grey. The peptide shown is human amyloid beta peptide, Aβ40.

PASTA energies and profiles. Two R scripts are provided to generate the same PDF graphics used in the web server. A simple shell script guides the overall program flow and a test sequence is also included. Details concerning installation and usage are explained in a 'README' file.

Acknowledgements

The authors wish to thank Fabrizio Chiti and Amos Maritan for ongoing collaboration on protein aggregation and Micky Del Favero for expert system administration.

Funding

This work was supported by Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale, grant 2005027330 in 2005. S.T. is funded by a 'Rientro dei cervelli' grant from the Italian Ministry for Education, University and Research (MIUR).

References

Caflisch,A. (2006) *Curr. Opin. Chem. Biol.*, **10**, 437–444.
Chiti,F. and Dobson,C.M. (2006) *Annu. Rev. Biochem.*, **75**, 333–366.
Chiti,F., Stefani,M., Taddei,N., Ramponi,G. and Dobson,C.M. (2003) *Nature*, **424**, 805–808.
Ferguson,N., *et al*. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 16248–16253.
Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) *Nat. Biotechnol.*, **22**, 1302–1306.
Fowler,D.M., Koulov,A.V., Alory-Jost,C., Marks,M.S., Balch,W.E. and Kelly,J.W. (2006) *PLoS Biol.*, **4**, e6.
Hoang,T.X., Marsella,L., Trovato,A., Seno,F., Banavar,J.R. and Maritan,A. (2006) *Proc. Natl Acad. Sci. USA*, **103**, 6883–6888.
Makin,O.S., Atkins,E., Sikorski,P., Johansson,J. and Serpell,L.C. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 315–320.
Petkova,A.T., Ishii,Y., Balbach,J.J., Antzutkin,O.N., Leapman,R.D., Delaglio,F. and Tycko,R. (2002) *Proc. Natl Acad. Sci. USA*, **99**, 16742–16747.
Sunde,M. and Blake,C. (1997) *Adv. Protein Chem.*, **50**, 123–159.
Trovato,A., Chiti,F., Maritan,A. and Seno,F. (2006) *PLoS Comput. Biol.*, **2**, 1608–1618.
Trovato,A., Maritan,A. and Seno,F. (2007) *J. Phys.: Condens. Matter*, **19**, 285221.

Received May 14, 2007; revised July 4, 2007; accepted July 6, 2007

Edited by Regina Murphy