Published online 25 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W469–W473***
doi:10.1093/nar/gkp351

NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation

Changsik Kim[1], Jiwon Choi[1], Seong Joon Lee[1], William J. Welsh[2] and Sukjoon Yoon[1,*]

[1]Sookmyung Women's University, Department of Biological Sciences, Hyochangwon-gil 52, Yongsan-gu, Seoul, Republic of Korea, 140-742 and [2]University of Medicine & Dentistry of New Jersey (UMDNJ), Department of Pharmacology, Robert Wood Johnson Medical School and the Informatics Institute of UMDNJ, 675 Hoes Lane, Piscataway, NJ 08854, USA

Received January 22, 2009; Revised April 15, 2009; Accepted April 22, 2009

ABSTRACT

The calculation of contact-dependent secondary structure propensity (CSSP) is a unique and sensitive method that detects non-native secondary structure propensities in protein sequences. This method has applications in predicting local conformational change, which typically is observed in core sequences of protein aggregation and amyloid fibril formation. NetCSSP implements the latest version of the CSSP algorithm and provides a Flash chart-based graphic interface that enables an interactive calculation of CSSP values for any user-selected regions in a given protein sequence. This feature also can quantitatively estimate the mutational effect on changes in native or non-native secondary structural propensities in local sequences. In addition, this web tool provides precalculated non-native secondary structure propensities for over 1 400 000 fragments that are seven-residues long, collected from PDB structures. They are searchable for chameleon subsequences that can serve as the core of amyloid fibril formation. The NetCSSP web tool is available at http://cssp2.sookmyung.ac.kr/.

INTRODUCTION

The sequence potential for non-native β-strand formation and the presence of chameleon sequences have been investigated extensively from the perspective that such structural features are implicated in the induction of fatal amyloid-related diseases (1–3). Our previous studies have shown that the α-helix and β-strand share similar sequence contexts and that the tertiary interaction is an important determinant of local secondary structure formation (4,5). Conventional secondary structure prediction methods, however, rely heavily on the intrinsic propensity of local sequences (6,7), and consequently they are not sensitive enough to predict non-native secondary structure formation. Thus, we have developed a computational method that quantifies the influence of tertiary interaction on secondary structural preference (4). Artificial neural network (ANN)-based algorithms that use preparameterized tertiary interactions with sequence inputs from users are designed to predict contact-dependent secondary structure propensities (CSSPs) (5,8).

Many attempts have been made to predict the aggregation-prone or amyloidogenic regions in protein sequences. The role of the physico-chemical properties of amino acids was investigated in determining the aggregation rate of a given sequence (9–11), and an optimal combination of physico-chemical properties of its amino acids provided a predictor, Zyggregator (12). Aggregation-prone fragments of amino-acid sequence were also predicted by using a statistical mechanics algorithm, TANGO (13). More recently, Trovato and his colleagues developed residue-based potentials to form parallel- or anti-parallel beta structure and used them to predict the core of amyloids (14,15). Structural features of core amyloid (16) have been also considered to evaluate amyloid fibril formation. The aggregation propensity of an inserted amino acid in the middle of the β-amyloid sequence has been experimental investigated and parameterized to predict the amyloidogenic propensities of other peptides (17). Despite these efforts, however, no rapid sequence-based methods have been reported to predict non-native secondary structure propensity in a globular protein and to pinpoint the aggregation-prone, core sequences of amyloid fibril formation. Thus, CSSP algorithms were proposed to evaluate the secondary structure propensity of a local sub-sequence in terms of tertiary interaction energies (4,5,8).

The CSSP methods, which adopt a fast machine learning algorithm, allow non-native secondary structure propensity in local sequences to be systematically evaluated with a step-wise increase of tertiary interaction energies. The trained single ANN exhibits 74% accuracy in predicting the native secondary structure of test sequences

*To whom correspondence should be addressed. Tel: +82 2 710 9415; Fax: +82 2 2077 7322; Email: yoonsj@sookmyung.ac.kr

in their native tertiary interaction energy state, and the dual ANN-based predictor has an 83% accuracy (440 884 SCOP20 fragments were used for the training, while 22 707 exclusive fragments from unique fold proteins were used for the tests) (8). In order to investigate the ability of NetCSSP in predicting amyloid fibril formation, we also retrieved two test sets of amino-acid fragments with experimental aggregation data from literature (13), and tested the output of single ANN for the predictability on aggregation-prone sequences (Figure 1). CSSP methods predict the secondary structure propensity for the center residue in a seven-residue sequence. Therefore we selected fragments of ≥10 amino acids length from the original test sets in order to obtain CSSP values for at least four residues in the middle. The sequence potential for aggregation was calculated from CSSP-derived P(helix), P(β) and P(coil) values in the form, ln(P(β)/[P(helix) × P(coil)] (5). The ROC plots in Figure 1 represent the sensitivity and specificity of the CSSP method in predicting the aggregation-prone sequences. Considering that the single ANN exhibits 74% accuracy in predicting secondary structures, the observed accuracy for predicting aggregation-prone fragments (i.e. AUC of 0.77 and 0.88 for test set 1 and 2) indicate that the CSSP provides an effective measure of the aggregation propensity. In our previous studies, we have already shown that CSSP methods can pinpoint the core of amyloid fibrils in many sequences (4,5,8). In addition, calculated CSSPs were shown to have a quantitative correlation with the aggregation rate of test fragments (5). All of these validation data are present in the 'Intro' page of the NetCSSP website.

We integrated these single and dual ANN methods into the current NetCSSP with a user-friendly web

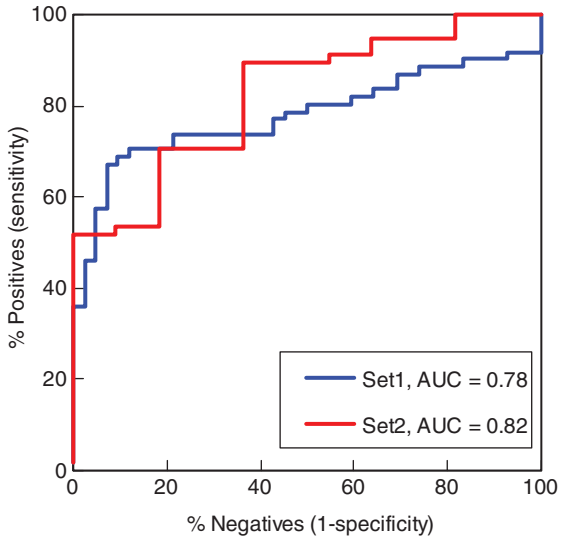

Figure 1. ROC plot validation of NetCSSP algorithm on two data sets. The sequence potential for aggregation was calculated from CSSP-derived P(helix), P(β) and P(coil) values in the form, ln(P(β)/[P(helix) × P(coil)] (5). Test set1 includes a total of 104 fragments of ≥10 amino-acid length, and test set2 includes 70 fragments of 10 amino-acid length. Both test sets were retrieved from literature (13). ROC plots represent the prioritization of aggregation-prone sequences over non-aggregates based on the CSSP values. AUC (Area Under the Curve) is in the range of 0–1 and represents the predictive power of the method.

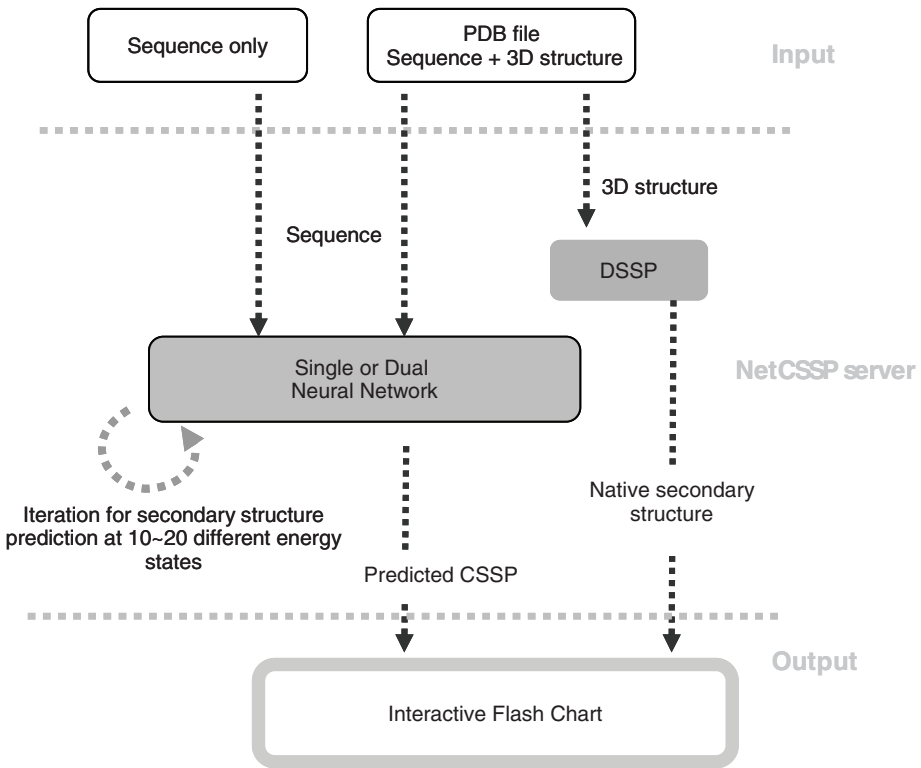

Figure 2. Workflow of CSSP calculation in the NetCSSP web server. Only sequence information is required for the CSSP calculation. When the 3D structure (in PDB format) is submitted, the predicted CSSP will be displayed in comparison with the native secondary structure information.

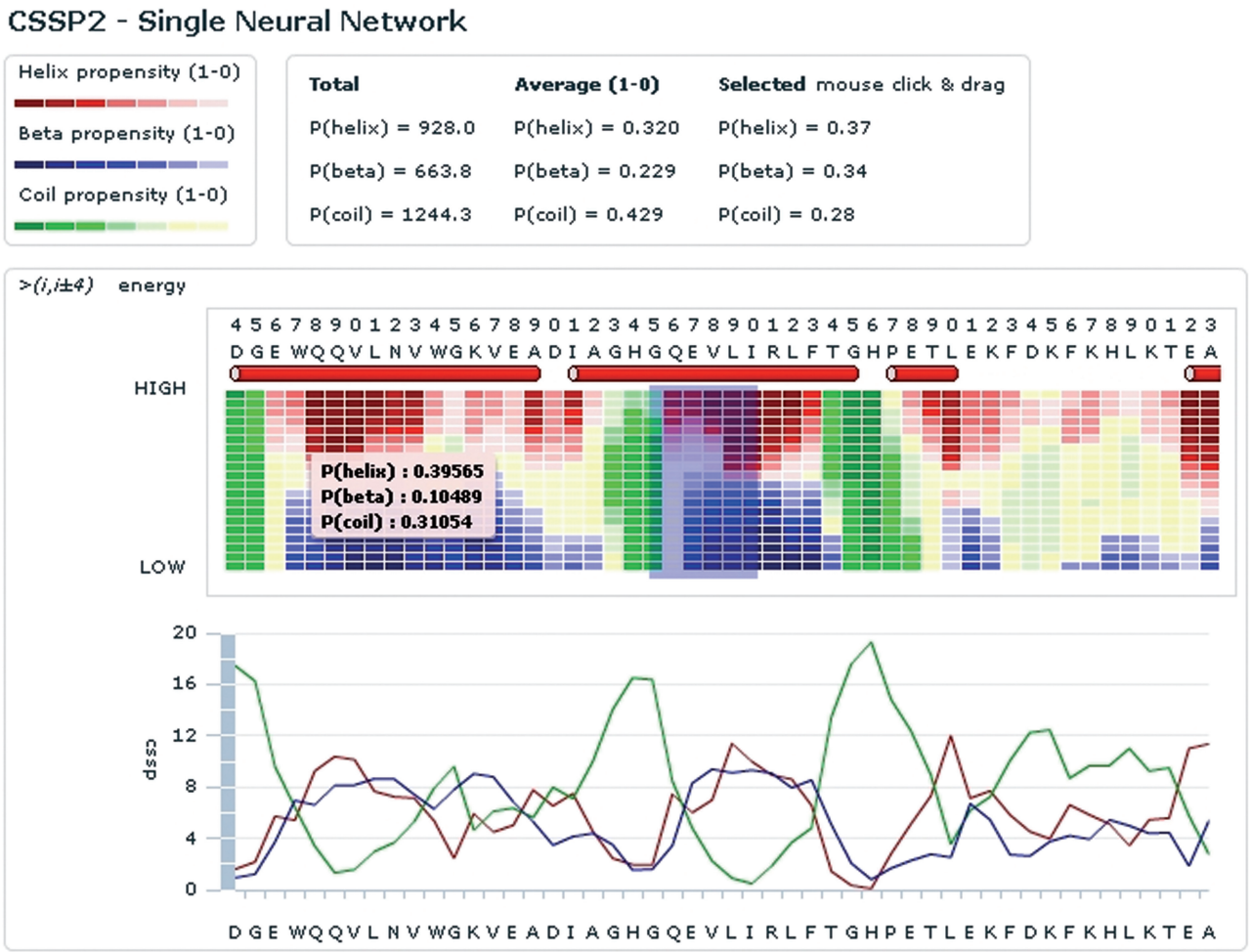

Figure 3. Output of single ANN mode NetCSSP profile of horse myoglobin (PDB ID: 1DWR). Only the N-terminal region (sequence 4–53) is displayed. The native helical conformation is displayed in red bars. The CSSP is predicted at 20 different energy steps of $>(i, I\pm4)$ interaction for helical, beta and coil propensities. The bottom diagram shows the sum of energy step-wise CSSPs. The additive CSSP values for the entire sequence and the residue-average values are given in the upper panel. One can also interactively calculate the CSSPs for any user-specified residues and energy steps. The light pink box shows the CSSP values for seventh residue, W, at an intermediate $>(i, I\pm4)$ energy level. The blue-shaded region represents a selection of 25-GQEVLI-30 sub-sequence and its CSSPs are presented at the upper panel.

interface (Figure 2). Because it returns CSSP profiles quickly, NetCSSP can be used for very long sequences or various combinations of amino-acid substitutions at particular sites. The easy Flash chart-based interface enables the interactive calculation of CSSP values for any user-selected regions in a given protein sequence. In addition, it compares experimental native secondary structures and predicted CSSPs when a PDB structure is inputted to the server. A third-party validation also has been reported and demonstrates that the CSSP calculation uniquely reveals local changes in β-strand propensity by mutations (18). This web tool also provides precalculated CSSPs with native secondary structure information for over 1 400 000 fragments that are seven-residues long, collected from SCOP90 domains. It is searchable for comparative evaluation of native and non-native secondary structure propensities and thus predicts amyloidogenic or chameleon sequences. We believe that the current NetCSSP is a unique tool for systematically predicting the sequence potential for local secondary structural changes. It has applications in protein engineering in addition to studies of amyloid fibrils.

NetCSSP PROFILE

NetCSSP provides a simple user interface to load input sequences. When a 3D structure file is loaded (in PDB format), the server automatically extracts sequence information for CSSP calculation and also runs the DSSP (Dictionary of Secondary Structure in Protein) program (19) to define native secondary structures (Figure 2). For CSSP calculation, the selected ANN runs multiple times with stepwise increases in preparameterized tertiary interaction energies [see ref. (8) for the detail].

A typical output of the single network-based calculation for a 3D structure input file of horse myoglobin (PDB ID: 1DWRa) shows the CSSP profiles and native secondary structure together (Figure 3). It provides the residue-based profile of secondary structure propensities in diverse tertiary interaction energies. Thus users can intuitively

Table 1. Search of chameleon sequences

Sequence	Secondary structure	PDB	Chain	SCOP	CSSP[a] (for native structure)	Non-native P(helix)	Non-native P(β)
GQEVLLT	CCEEEEE	1o89	A	b.35.1.2	0.48	0.3	–
QEVLLVQ	HHHHHHH	1a8o	–	a.28.3.1	0.43	–	0.33
QEVLLWL	HHHHHHH	1csh	–	a.103.1.1	0.46	–	0.36
TLAQEVL	HHHHHHH	1e1o	A	d.104.1.1	0.52	–	0.23
AQEVLLA	EEEEEEE	1exs	A	b.60.1.1	0.28	0.51	–
KPIQEVL	CCHHHHH	1het	A	c.2.1.1	0.55	–	0.21
QEVLKSI	HHHHHHH	1mg7	A	d.14.1.6	0.53	–	0.22
NLQEVLG	CCCEEEC	1n3l	A	c.26.1.1	0.33	0.41	–
LQEVLNT	HHHHHHH	1odf	A	c.37.1.6	0.55	–	0.24
QEVLLPR	CEEEECC	1ojq	A	d.166.1.1	0.51	0.19	–
AHQEVLF	EEEEEEE	1p9l	A	d.81.1.3	0.31	0.34	–
IQEVLEV	HHHHHCC	1qgu	B	c.92.2.3	0.53	–	0.34
QEVLETM	HHHHHHH	1tml	–	c.6.1.1	0.58	–	0.27

The subsequence (GQEVLI) in the shaped box in Figure 3 has both strong helical and beta propensities. Searching the fragment database, including precalculated CSSPs values, shows that GQEVL and QEVL are found in both helical and beta contexts in various native proteins. The native secondary structure is represented by C (coil), E (extended β) and H (helix).
[a]CSSP represents the calculated propensity for the native secondary structure for a seven-residue sequence. For example, when a residue adopts 'coil' for the native structure, P(coil) of calculated CSSPs was selected.

identify the potential amyloidogenic subsequences from the CSSP profile. In the present display, the entire myoglobin sequence adopts a primarily helical conformation in the native structure. The myoglobin sequence has been reported to form amyloid fibrils by switching its helices to β-strand conformations (2). In particular, the N-terminal region (1–29) is known to have a high propensity for β-aggregation (20). The CSSP profile in Figure 3 shows high helical and beta propensities that are consistent with the previous experimental observation. The propensity for each of three secondary structure elements, helix, β-strand and coil, is calculated at 20 different levels of $>(i, I \pm 4)$ interaction energy. Most of the N-terminal regions show both strong native helical propensity at high $>(i, I \pm 4)$ interaction energies and non-native beta propensity at low $>(i, I \pm 4)$ interaction energies.

Users can also quantitatively analyze the CSSP profile interactively. For example, Figure 3 shows a selection, 25-GQEVLI-30, which is included in the highly amyloidogenic N-terminal sequence of horse myoglobin. NetCSSP returns detailed CSSP values for the selected sequence at the upper panel. GQEVLI shows similar propensity to form a helix and β-strand [i.e. P(helix) = 0.37 and P(β) = 0.34], although the entire myloglobin sequence shows a higher helical propensity than the β-strand propensity [i.e. P(helix) = 0.320, P(β) = 0.229]. The diagram at the bottom of Figure 3 shows the residue-based sum of CSSPs, which clearly shows that GQEVLI is a potential hot spot for accelerating amyloid fibril formation.

CHAMELEON SEQUENCES

Non-native secondary structure propensities that are predicted from CSSP profiles can be directly confirmed by searching chameleon sequences, which are also provided by the present NetCSSP server. For example, the search output for the GQEVLI query that is selected in Figure 3 shows that GQEVL and QEVL are found in both helical and beta contexts in the native structures of various proteins (Table 1). This search result implies that the GQEVLI sequence in myoglobin can form non-native beta conformations by altering tertiary interactions during the course of amyloid fibril formation.

This chameleon sequence database includes 1 424 079 seven-residue fragments that are extracted from 2339 unique fold SCOP20 domains. The native 3D structure of each fragment was obtained from the PDB structure, and the CSSP was calculated within the context of a complete 3D structure of the proteins. User-defined query sequences can include up to seven residues. By searching this database, you can directly compare the experimental native secondary structures with the calculated CSSPs for native and non-native secondary structures. Table 1 shows that GQEVL and QEVL are found in both helical and beta contexts in the native structure. Consistently, the calculated CSSPs show that they have similar propensity for native and non-native secondary structures in many different protein contexts. One also can search the database using a cutoff for CSSP values. Table 2 shows search outputs of the top-five lists for the highest non-native helical and beta propensities when the database is searched using cutoffs for non-native P(helix) and non-native P(β). Complete information on the seven-residue sequences, PDB ID, SCOP ID, native secondary structures, and CSSP values are available. The information in the chameleon database is useful for designing new ambivalent or non-ambivalent peptide sequences, as well as identifying amyloidogenic chameleon subsequences in a given protein.

FUNDING

Basic Research Program of the Korea Science & Engineering Foundation [grant No. R01-2006-000-10515-0]; SRC program of MOST/KOSEF (Research Center for Women's Diseases); Korea Research Foundation Grant funded by the Korean Government

Table 2. Output of search of chameleon sequences with the highest non-native P(helix) and non-native P(β) values

Sequence	Secondary structure	PDB	Chain	SCOP	CSSP (for native structure)	Non-native P(helix)	Non-native P(β)	Relative P(helix)	Relative P(β)
LRRARAA	CCCCCCC	1cer	O	d.81.1.1	0.18	0.69	0.13	3.93	0.73
KQMLAKA	CCCCCCC	1goj	A	c.37.1.9	0.13	0.68	0.18	5.10	1.33
QEQLEKA	CCCCCCC	1gx5	A	e.8.1.4	0.17	0.68	0.13	3.89	0.77
AKEAAQK	CCCCCCC	1g9l	A	a.144.1.1	0.2	0.68	0.1	3.50	0.53
ARAQARQ	CCCEEEE	1omh	A	d.89.1.5	0.14	0.68	–	4.77	–
AVIVVFD	CCCCCCC	1bgx	T	c.120.1.2	0.15	0.22	0.58	1.45	3.80
VTVTVFD	CCCCCCC	1eu1	A	b.52.2.2	0.3	0.12	0.57	0.41	1.88
VFEVNIR	HHHHHHH	1nxc	A	a.102.2.1	0.23	–	0.57	–	2.52
VYWFTVE	HHHCCCC	1toh	–	d.178.1.1	0.22	–	0.57	–	2.54
VYVVFSV	CCCCCCC	1vho	A	c.56.5.4	0.16	0.23	0.57	1.45	3.52

By searching the fragment DB, one can quantitatively analyze non-native secondary structure propensities in comparison with native secondary structure patterns.

(MOEHRD) [KRF-2006-311-C00582]; grant from the KRIBB Research Initiative Program. Funding for open access charge: SRC program of MOST/KOSEF (Research Center for Women's Diseases).

Conflict of interest statement. None declared.

REFERENCES

1. Sacchettini,J.C. and Kelly,J.W. (2002) Therapeutic strategies for human amyloid diseases. *Nat. Rev. Drug Discov.*, **1**, 267–275.
2. Fandrich,M., Fletcher,M.A. and Dobson,C.M. (2001) Amyloid fibrils from muscle myoglobin. *Nature*, **410**, 165–166.
3. Yoon,S. and Jung,H. (2006) Analysis of chameleon sequences by energy decomposition on a pairwise per-residue basis. *Protein J.*, **25**, 361–368.
4. Yoon,S. and Welsh,W.J. (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci.*, **13**, 2149–2160.
5. Yoon,S. and Welsh,W.J. (2005) Rapid assessment of contact-dependent secondary structure propensity: relevance to amyloidogenic sequences. *Proteins*, **60**, 110–117.
6. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, **266**, 525–539.
7. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
8. Yoon,S., Welsh,W.J., Jung,H. and Yoo,Y.D. (2007) CSSP2: an improved method for predicting contact-dependent secondary structure propensity. *Comput. Biol. Chem.*, **31**, 373–377.
9. DuBay,K.F., Pawar,A.P., Chiti,F., Zurdo,J., Dobson,C.M. and Vendruscolo,M. (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.*, **341**, 1317–1326.
10. Tartaglia,G.G., Cavalli,A., Pellarin,R. and Caflisch,A. (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci.*, **14**, 2723–2734.
11. Chiti,F., Stefani,M., Taddei,N., Ramponi,G. and Dobson,C.M. (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, **424**, 805–808.
12. Tartaglia,G.G. and Vendruscolo,M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev.*, **37**, 1395–1401.
13. Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
14. Trovato,A., Chiti,F., Maritan,A. and Seno,F. (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput. Biol.*, **2**, e170.
15. Trovato,A., Seno,F. and Tosatto,S.C. (2007) The PASTA server for protein aggregation prediction. *Protein Eng. Des. Sel.*, **20**, 521–523.
16. Thompson,M.J., Sievers,S.A., Karanicolas,J., Ivanova,M.I., Baker,D. and Eisenberg,D. (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc. Natl Acad. Sci. USA*, **103**, 4074–4078.
17. Conchillo-Sole,O., de Groot,N.S., Aviles,F.X., Vendrell,J., Daura,X. and Ventura,S. (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
18. Higashimoto,Y., Asanomi,Y., Takakusagi,S., Lewis,M.S., Uosaki,K., Durell,S.R., Anderson,C.W., Appella,E. and Sakaguchi,K. (2006) Unfolding, aggregation, and amyloid formation by the tetramerization domain from mutant p53 associated with lung cancer. *Biochemistry*, **45**, 1608–1619.
19. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
20. Monsellier,E., Ramazzotti,M., de Laureto,P.P., Tartaglia,G.G., Taddei,N., Fontana,A., Vendruscolo,M. and Chiti,F. (2007) The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution. *Biophys. J.*, **93**, 4382–4391.