Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 31 (2007) 373–377
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Brief Communication

CSSP2: An improved method for predicting contact-dependent secondary structure propensity

Sukjoon Yoon [a,*], William J. Welsh [b], Heeyoung Jung [a], Young Do Yoo [c]

[a] *Sookmyung Women's University, Department of Biological Sciences, Research Center for Women's Diseases (RCWD), Hyochangwongil 52, Yongsan-gu, Seoul 140-742, Republic of Korea*
[b] *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*
[c] *Korea University, College of Medicine, Graduate School of Medicine, Anam-dong 126-1, Sungbuk-ku, Seoul 136-705, Republic of Korea*

Received 29 March 2007; received in revised form 3 June 2007; accepted 4 June 2007

Abstract

The calculation of contact-dependent secondary structure propensity (CSSP) has been reported to sensitively detect non-native β-strand propensities in the core sequences of amyloidogenic proteins. Here we describe a noble energy-based CSSP method implemented on dual artificial neural networks that rapidly and accurately estimate the potential for the non-native secondary structure formation in local regions of protein sequences. In this method, we attempted to quantify long-range interaction patterns in diverse secondary structures by potential energy calculations and decomposition on a pairwise per-residue basis. The calculated energy parameters and seven-residue sequence information were used as inputs for artificial neural networks (ANNs) to predict sequence potential for secondary structure conversion. The trained single ANN using the $>(i, i \pm 4)$ interaction energy parameter exhibited 74% accuracy in predicting the secondary structure of test sequences in their native energy state, while the dual ANN-based predictor using $(i, i \pm 4)$ and $>(i, i \pm 4)$ interaction energies showed 83% prediction accuracy. The present method provides a simple and accurate tool for predicting sequence potential for secondary structure conversions without using 3D structural information.

Keywords: Amyloid fibril formation; Secondary structure prediction; Machine learning; Artificial neural network; Energy decomposition

1. Introduction

The conversion of α-helix to β-strand conformations and the presence of chameleon sequences have been widely reported and extensively investigated from the perspective that such structural features are implicated in the induction of amyloid-related fatal diseases (Chiti et al., 1999; Jimenez et al., 1999; Fandrich et al., 2001; Sacchettini and Kelly, 2002). Previous studies have shown that the propensity of individual amino acids to adopt particular secondary structures arises from a combination of local factors (inherent conformational preferences) and non-local factors (tertiary effects) (Minor and Kim, 1996; Sudarsanam, 1998). However, conventional secondary structure prediction methods rely heavily on intrinsic propensity and local neighbors (Rost, 1996; Pollastri et al., 2002). Thus, we have recently introduced a computational method that quantifies the influence of tertiary effects on secondary structural preferences using a simple approach that counts the number of atom-to-atom tertiary contacts (TCs) (Yoon and Welsh, 2004, 2005). Employing this TC-based scheme, we formulated a computational tool that predicts contact-dependent secondary structure propensity (CSSP). Sequence–structure relationships of query sequences were systematically evaluated in terms of TCs (low TCs versus high TCs) by analyzing the secondary structure preferences of template sequences for which the three-dimensional structure is known. Accurate predictions of non-native secondary structure preferences were obtained using short (seven-residue) query sequences without direct knowledge of the query's native tertiary structure and despite the absence of structural information on amyloidogenic sequences.

In the present study, we attempted to improve the CSSP method by using energy-based tertiary interaction parameters as inputs rather than the simple TC counting. We recently analyzed the observed secondary structure conversion in chameleon

* Corresponding author.
E-mail addresses: yoonsj@sookmyung.ac.kr (S. Yoon), welshwj@umdnj.edu (W.J. Welsh), skylooker@sookmyung.ac.kr (H. Jung), ydy@kumc.or.kr (Y.D. Yoo).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.06.002

sequences by using potential energy decomposition on a pairwise per-residue basis (Yoon and Jung, 2006). The long range interaction beyond the $(i, i \pm 4)$ position, i.e., $>(i, i \pm 4)$ interaction energy, was found to be an effective tertiary interaction parameter for discriminating beta conformation from helix or random coil conformations in chameleon sequences. The $(i, i \pm 4)$ term was more discriminative than $>(i, i \pm 4)$ energy for the helical conformation of chameleon sequences. In addition, electrostatic and polar solvation terms were shown to be major energetic factors in secondary structure conversion in chameleon sequences. In the present study, we investigated how these energetic parameters could improve the CSSP method compared with TC-based method. Our primary motivation in developing an advanced CSSP algorithm was to provide a simple yet accurate tool that can gauge the non-native secondary structure propensity and the marginal stability of local sequences.

2. Methods

2.1. Preparation of Peptide Library

To construct the peptide library of a seven-residue sequence from diverse tertiary contexts, we used protein domain sequences and their 3D structures listed in Structural Classification Of Proteins 20 (SCOP20) Version 1.67 (Brenner et al., 2000). SCOP20, a collection of protein domains that exhibit <20% sequence identity between any two members, provided a rich source of non-homologous sequence contexts from diverse tertiary environments. A total of 3676 globular domains whose 3D structures have been determined by X-ray crystallography were selected from SCOP20 after excluding membrane proteins, small proteins, and proteins with incomplete structural information. The 3D coordinates of these domains were retrieved from the Protein Data Bank (PDB). In order to optimize the protein structures for energetic analysis, energy minimization was carried out on the retrieved PDB entries using the AMBER program (Version 8.0) (Case et al., 2004). After energy minimization, the secondary structure of each protein residue was assigned by using the dictionary of secondary structure in proteins (DSSP) program (Kabsch and Sander, 1983) which is one of popular methods (such as P-SEA) for identifying secondary structures in protein structures. Then, a sliding seven-residue window, shifting one residue at a time, was employed to collect a total of 463,591 sequence fragments of seven-residue length from the 3676 SCOP20 domain structures.

2.2. Pairwise Per-residue Energy Calculation and Energy Decomposition

For the center residue in each seven-residue sequence in a given protein, interaction energies with the remaining residues were calculated on a per-residue basis, particularly for two interaction types, i.e., $(i, i \pm 4)$ and $>(i, i \pm 4)$ interactions. The energy for $>(i, i \pm 4)$ interactions represents the sum of pairwise potential energies with residues beyond the $(i, i \pm 4)$ positions in a sequence. The solvation effect was represented implicitly by a generalized Born/surface area (GB) model implemented in the Sander module of AMBER. Residue-based non-bonding interaction energies were further decomposed into van der Waals and electrostatic terms. The solvation energy was also calculated and decomposed into polar and solvent accessible surface (SAS) components of the GB model.

Since individual amino acids differ with respect to side-chain length, composition, and hydrophobicity, the interaction energy of an input sequence was standardized by the average and standard deviation of interaction energy of the corresponding amino acids. Then, the standardized potential energy for a seven-residue fragment was obtained by summing the individual energies of the middle five residues. The TC counting in minimized structures was performed in the same fashion to a previous study (Yoon and Welsh, 2005).

2.3. Artificial Neural Network for CSSP Prediction

A feed-forward back-propagation artificial neural network (ANN) was implemented using the Stuttgart Neural Network Simulator (SNNS Version 4.1, http://www-ra.informatik.uni-tuebingen.de/SNNS/). The single network architecture consists of perceptrons with an input layer encoding for the seven-residue window, a single hidden layer with 24 nodes, and a single output layer comprising three output nodes corresponding to α-helix, β-strand and random coil. The dual network architecture consists of the two single networks described above except that two distinct output nodes were used. One network has output nodes for α-helix and non-helix, thus predicting helical propensity; the other network has output nodes for β-strand and non-beta strand, thus predicting beta propensity. The input layer contains an additional variable node for the query sequence. This variable node was set to the computed energy term ($>(i, i \pm 4)$ energy) or TC value of the query sequence for single network architecture during the training and testing phases. However, for the dual network architecture, the α-helix predicting network uses the $(i, i \pm 4)$ energy term of the query sequence as the additional input, while the beta predicting network uses $>(i, i \pm 4)$ energy term as the additional input. The training set comprised 440,884 SCOP20 fragments, each seven residues in length, together with their energies or TC values, while the test set consisted of 22,707 fragments that were extracted from 1629 unique fold domains to evaluate the predictive performance of the trained ANNs. The prediction accuracy of the trained ANNs was measured using the standard Q3 score, which is the sum of correct predictions divided by the total number of test queries.

2.4. Measure of Secondary Structure Propensity on Protein Sequence

The alpha propensity, $P(\alpha)$, and beta propensity, $P(\beta)$, of a residue (or fragment) were calculated from the trained ANNs. In the case of dual network architecture, $P(\alpha)$ and $P(\beta)$ were calculated from the two separate networks independently. $P(\alpha)$ is the sum of output values of the helix node from the helix prediction network using the $(i, i \pm 4)$ energy values ranging from -2.0 to 2.0. $P(\beta)$ is the sum of output values of beta node