Analysis of the secondary structure of a protein’s N-terminal

C G Floare¹, M Bogdan¹, O Horovitz², A Mocanu² and M Tomoaia-Cotisel²

¹ National Institute for R & D of Isotopic and Molecular Technologies, 65-103 Donath, 400293 Cluj-Napoca, Romania
² Babes-Bolyai University, Faculty of Chemistry and Chemical Engineering, Department of Physical Chemistry, 11 Arany Janos, 400028 Cluj-Napoca, Romania

E-mail: calin.floare@itim-cj.ro

Abstract. The major protein component from aleurone cells of barley (Hordeum vulgare L.), PACB, is related to 7S globulins present in other cereals and to the vicilin-type 7S globulins of legumes and cotton seed. It contains 4 subunits of about 20, 25, 40 and 50 kDa molecular weights. The N-terminal sequence of 16 amino acids (over 260 atoms) in the protein was previously determined, and our aim is the prediction of its secondary structure. The empirical Chou-Fasman method was applied in an improved version as well as the empirical DSC method (discrimination of protein secondary structure class) with quite similar results. A molecular dynamics simulation was also performed, using the FF99SB forcefield within AMBER version 9.0. Solvation effects were incorporated using the Born model. The results are compared and a 3D model is proposed.

1. Introduction

Proteins adopt 3-dimensional structures, which enable them to function in specific ways, for example as biologically active enzymes. Shorter polypeptides generally adopt less compact and less stable structures in solution, making the experimental characterization of their preferred folds a non-trivial task. The end of the polypeptide chain with an unbound amine group, the N-terminus or N-terminal, is usually the first part of the protein that exits the ribosome during protein biosynthesis. It often contains sequences that act as targeting signals, basically intracellular zip codes that allow for the targeting of the protein to its designated location within the cell. Therefore, there is a growing interest in the characterization of these N-terminal signal sequences and of their role [1, 2, 3]. Plants store amino acids for longer periods in the form of specific storage proteins, deposited in seeds, in root and shoot tubers or in other vegetative organs [4]. The latest advances in computer technology and modelling techniques facilitate simulations of small peptide folding in atomic detail. In this work, we are trying to predict the secondary structure for the N-terminal of the major globular storage protein from aleurone cells of barley (Hordeum vulgare L.). This protein is related to 7S globulins present in other cereals and to the vicilin-type 7S globulins of legumes and cotton seed. It contains 4 subunits of about 20, 25, 40 and 50 kDa molecular weights. Its N-terminal sequence of 16 amino acids in the protein is given as follows: X²Glu³Gln⁴Gly⁵Asp⁶Ser⁷Arg⁸Arg⁹Pro¹⁰Tyr¹¹Val¹²Phe¹³Gly¹⁴Pro¹⁵Arg¹⁶(Ser or His)¹⁷Phe, or with the one-letter notations of amino-acids XEQGDSRPPYVGPRSHF¹⁷,
where X stands for the first amino acid of the N-terminal of aleurone protein which was not identified [5]. The secondary structure of this protein was investigated by advanced spectroscopies [6, 7, 8, 9] and its adsorption at various solid and liquid interfaces and on nanoparticles was studied [10, 11].

2. Methodology
We started by using two empirical techniques. The Chou-Fasman method is used to predict secondary structure in proteins and depends on assigning a set of prediction values to a residue and then applying a simple algorithm to those numbers [12]. The method is based on analyses of the relative frequencies of each amino acid in α-helices, β-sheets, and turns based on known protein structures solved with X-ray crystallography. From these frequencies a set of probability parameters were derived for the occurrence of each amino acid in each secondary structure type, and these parameters are used to predict the probability that a given sequence of amino acids would give a certain conformation in a protein. The Chou-Fasman parameters found some strong tendencies among individual amino acids to prefer one type of secondary structure over others. Alanine, glutamate, leucine, and methionine were identified as helix formers, while proline and glycine commonly end a helix. The original Chou-Fasman parameters were derived from a very small and non-representative sample of protein structures due to the small number of such structures that were known at that time. These original parameters have since been shown to be unreliable and have been updated from a current dataset, along with modifications to the initial algorithm [13]. The DSC Method (Discrimination of protein Secondary structure Class), fully described by King & Sternberg [14], is accessible as ‘Secondary Structure Prediction’ (SSP) option from Accelrys Discovery Visualizer v1.7.

On the other hand, molecular dynamics (MD) is a method in which Newton’s equations of motion are integrated to reproduce the time evolution (trajectory) of the atoms in a system and is widely used to study dynamic processes in biomolecular systems at atomic resolution. MD simulations have been shown to be a reliable instrument to investigate the process of reversible folding of various peptides [15, 16]. Encouraging agreement with experiment has been obtained, showing that the biomolecular forcefields mimic microscopic processes very well. On a classical MD simulation of a biomolecular system some important points must be emphasized. Firstly, the dynamics of individual atoms during the process of folding cannot yet be directly observed experimentally. Experimental data at atomic resolution is only available for equilibrium distributions of conformations under specific conditions. Secondly, the time scale on which the relevant biomolecular systems can be simulated in atomic detail is currently on the order of picoseconds to nanoseconds, while the lower limit for proteins as well as most peptides to fold is believed to be microseconds. Thirdly, the process of peptide folding involves many (peptide and solvent) degrees of freedom and is governed by small (free) energy differences, on the order of a few multiples of kT (k - Boltzmann’s constant), which puts very high demands on the accuracy of the atomic interaction function and model that is used [17]. We specify also that when an experimentally determined structure is not available it is difficult to evaluate the quality of conformations sampled on the trajectory. Consequently, there is some inherent uncertainty in this kind of study at which we add the fact that we do not consider the interactions with the body of the protein (which hasn’t been determined).

3. Computational Details
The on-line Chou-Fasman predictor CHOFAS version 2.0u66 September 1998 was used for the two N-terminals of barley aleurone, one containing serine (S) and the other histidine (H) residue. SSP option in Accelrys Discovery Visualizer v1.7 (based on DSC method) was also used. The MD simulations have been performed using AMBER version 9.0 [17], starting with the linear structure: GLU GLN GLY ASP SER ARG ARG PRO TYR VAL PHE GLY PRO ARG SER
PHE, built using xLEaP tool. Initially a short minimization was performed (1000 steps - 500 using steepest descent followed by 500 of conjugate gradient algorithms) to cleanup the structure and fix up hydrogen positions, such that the MD be stable. We ran a non-periodic simulation and, since this system has only 16 residues, we could afford to use a large non-bonded cutoff of 999˚A and to include all atom pairs in the effective Born radii calculation. The MD was started using sander executable: the system was slowly heated up to 325K over 50 ps in a total of 7 stages. In this way, the chances that the system will blow up are reduced, by allowing it to equilibrate at each temperature. The temperature, slightly higher than the normal 300K, was chosen in order to avoid the system being kinetically trapped in local minima. The simulation was performed in the weak-coupling canonical ensemble, using Berendsen thermostat [18] with a time constant for heat bath coupling of 1 ps, the Stony Brook parameterization of the FF99 forcefield and the implicit Born solvation model [19], which includes the solvent effects within the forcefield equations. To allow the initial system to relax in a controlled fashion, a very short time step of 0.5 fs was used for the heating stage and then it was switched to 2 fs for the equilibration. On this stage, where there is an increased chance of system instability, we also set the coordinate write frequency on the trajectory to 50, in order to permit the identification of possible problems occurring. On the production MD, a more appropriate value of 500 was used. With this equilibrated structure, a long production run at 325 K was started: a 200 ns simulation using the same forcefield and the time step 2 fs. On account of this large time step, all bonds involving hydrogen had to be constrained. The SHAKE algorithm was used on the whole time. The Berendsen thermostat was used for temperature control, more closely coupled (system time constant for heat bath coupling of 0.5 ps) to keep the system close to 325 K.

4. Results and Discussion
The on-line Chou-Fasman predictor and the DSC method gave the following results for the N-terminal of barley aleurone containing S (serine) residue (the analysis of the sequence containing histidine residue gave essentially similar results):

| Chou-Fasman Predictor   | DSC Method     |
|-------------------------|----------------|
| EQGDSSRPYVFGPRSF        | EQGDSSRPYVFGPRSF |
| helix                   | DSC-SEC        |
|                         | CCCCCCCCEEEM  |
| sheet                   | PROB-H         |
|                         | 0001001000 011000 |
| turns                   | PROB-E         |
|                         | 11111114688 742111 |
|                         | PROB-C         |
|                         | 9998995422 357999 |

Table 1. The results from Chou-Fasman predictor are straightforward. There are turns at prolyl residues and sheet between them. The first section of the report for the DSC method shows 5 lines for each sequence: (1) The sequence; (2) The predicted secondary structure type (DSC_SEC); (3, 4, 5) The probability of being helix (PROB_H), a beta strand (PROB_E) or a coil (PROB_C). [Values of PROB_H, PROB_E, and PROB_C are from 0 to 9, with 9 being the highest probability.]

From Table 1 we can infer that we have a coil structure with turns where the prolyl residues are present and sheet between them. A database analysis of X-ray diffraction structures of 7S seed storage proteins reveals [20, 21] that these proteins are typically trimeric, lacking cysteine residues (it cannot form disulfide bonds). The β-sheet conformation predominates and the N- and C-terminals are generally coils. The MD computed trajectory for the N-terminal containing
the serine residue is analyzed by plotting temperature and potential energy from the output file. From figure 1 it is evident that the temperature rose smoothly and equilibrated at 325 K. Similarly, changes in the kinetic and potential energy should be smooth and should level off at equilibrated values. Any sudden spikes in the temperature or energy plots are indicative of a problem with our simulation protocol (bad starting structure, too long time step, inappropriate parameters etc.).

In figure 2A is shown the potential energy as a function of time during this simulation. The highest energy difference between conformations is 115.9 kcal/mol, but overall a decrease of around 50 kcal/mol can be observed. The potential energy plot helps to locate the structure with the lowest energy, considered to be the folded state. In this case it’s showing an initial important decrease over the course of the first 10 ns and then, in steps, the system explore different conformations until the moment when, before 160 ns, it reaches the minimum.

Using the ptraj program, the backbone root mean square deviation (RMSd) of structures relative to the lowest-energy conformation has been calculated and is represented in figure 2B. From the analysis of this representation and of the trajectory, using VMD and Sirius programs, it seems that the N-terminal of PACB stays most of the time in alternative structures and has not folded up to the native structure. Even if the trajectory is fairly long, the system is still exploring the potential energy surface, since the end point of the trajectory shows that the system tends to unfold and fold up differently. Here we must also keep in mind the fact, mentioned before, that shorter polypeptides generally adopt less compact and less stable structures in solution. This simulation should be continued, but we can consider the conformation having the minimum potential energy as an intermediary folded state.
The N-terminal structure corresponding to the lowest potential energy is given in Figure 3. It potential energy is -894.6279 kcal/mol, and the model is already well folded and presents hydrogen bonds between amino acids. In the model there are turns at prolyl residues, a fact which is consistent with the previous analysis, with the empirical methods. The Molecular graphics image was produced using the UCSF Chimera package [22].

5. Conclusions
In order to determine the secondary structure of the N-terminal of the major protein component present in embryos and aleurone layers of barley, we used two different approaches. Initially we analyzed this system using two empirical prediction methods. Additionally, we studied the folding of this peptide by performing a 200 ns all-atom fully unrestrained molecular dynamics simulation in implicit solvent. Both empirical methods predicted a coil conformation with two turns where the prolyl residues are present, and sheet conformation between these 2 turns. For the structure containing the serine residue, analyzed by molecular dynamics, the conformation found to have the lowest potential energy has a coil structure and presents turns at prolyl residues, i.e. consistent with the previous empirical results. Actually, in known structures of other 7S seed storage proteins, determined by X-ray diffraction, the $\beta$-sheet conformation predominates and the N- and C-terminals are generally coils. From the RMSd between the trajectory structures and lowest energy structure found in the simulation, it seems evident that the N-terminal of PACB spends most of time in alternative structures and has not yet folded up to the native structure, making further analysis necessary. This could be achieved in several ways. A longer trajectory could be run (possibly with the use of explicit solvent). Alternatively, an annealing approach or a replica exchange could be simulated.

References
[1] Meijer A B, Spruijt R B, Wolfs C J A M and Hemminga M A 2001 Biochemistry 40 5081
[2] Hegde R S and Bernstein H D 2006 Trends in Biochemical Sciences 31 563
[3] Meinnel T and Giglione C 2008 Proteomics 8 626
[4] Miintz K 1998 Plant Molecular Biology 38 77
[5] Yupsanis T, Burgess S R, Jackson P J and Shewry P R 1990 Journal of Experimental Botany 41 385
[6] Tomoaia-Cotisel M 2006 in Convergence of Micro-Nano-Biotechnologies, Series in Micro and Nanoengineering ed M Zaharescu, E Burzo et al (Bucharest: Romanian Academy Press) 9 147
[7] Tomoaia-Cotisel M, Tomoaia-Cotisel A, Yupsanis T, Tomoaia G, Balea I, Mocanu A and Racz C 2006 Revue Roumaine de Chimie 51 1181
[8] Tomoaia-Cotisel M, Mocanu A, Leopold N, Vasilescu M, Chis V and Cozar O 2007 JOAM 9 637
[9] Bratu I, Tomoaia-Cotisel M, Damian G and Mocanu A 2007 J. Optoelec. Adv. Mat. 9 672
[10] Horovitz O, Tomoaia G, Mocanu A, Yupsanis T and Tomoaia-Cotisel M 2007 Gold Bulletin 40 213 & 295
[11] Tomoaia-Cotisel M, Horovitz O, Borostean O, Bobos L D, Tomoaia G, Mocanu A and Yupsanis T 2008 Studia Universitatis Babes-Bolyai, Chemia 53 143
[12] Chou P Y and Fasman G D 1974 Biochemistry 13 211 & 222
[13] Chen H, Gu F and Huang Z 2006 BMC Bioinformatics 7 (Suppl 4) S14
[14] King R D and Sternberg M J E 1996 Protein Science 5 2298
[15] Shea J E and Brooks C L 2001 Annual Reviews of Physical Chemistry 52 499
[16] Onufriev A, Case D A, Cheatham T E, III, Darden T, Luo R and Simmerling C 2003 J. Comput. Chem. 24 1668
[17] Berendtsen H J C, Postma J P M, Gunsteren van W F, DiNola A and Haak J R 1984 Journal of Chemical Physics 81 3684
[18] Tsui V and Case D A 2001 Biopolymers (Nucleic Acid Sciences) 56 275
[19] Ko T P, Ng J D, Greenwood A and McPherson A 1993 Acta Cryst. D. Biol. Cryst. 49 478
[20] Maruyama N, Adachi H, Takahashi K, Yagasaki K, Kohno M, Takenaka Y, Okuda E, Nakagawa S, Mikami B and Utsumi S 2001 European Journal of Biochemistry 268 3595
[21] Pettersen E F, Goddard T D, Huang C C, Couch G S, Greenblatt D M, Meng E C and Ferrin T E 2004 Journal of Computational Chemistry 26 900