Prediction of α-helix Using Data Mining Decision Tree Technique

Mazoni, I.¹; Salim, J.A.¹; Yano, I.H.¹; Moraes, F.R.¹; Carvalho, J.G.¹; Jardine, J.G.¹; Neshich, I.¹; Neshich, G.¹

¹Embrapa Informática Agropecuária

The primary sequences of all-α proteins were extracted from the StingRDB. These sequences were aligned based on the occurrence of the secondary structure elements (SSE) of the size $n$, with $n$ ranging from 5 to 12 amino acids (AA). We used only the sequences with the same number of amino acids belonging to a SSE, having the same size and having a consensus definition of the type of SSE in three algorithms/identifiers: PDB, DSSP and Stride. To each found sequence another 32 residues were added: before and after the selected SSE. For each amino acid constituting the SSE 28 structure/function/sequence descriptors/parameters were extracted from the STING_DB, covering information in terms of amino acid conservation, contacts established with other AA, geometry, structure and physicochemical of its environment. The descriptors/parameters used were Distance from N_Term, Eletrostatic_Potential@$C$α, Hydrophobicity_Isolation, Dihedral_Angle_ψ, Dihedral_Angle_φ, Cross_Pres_Order_Cα, Cross_Link(Order_Cα, Relative_Entropy_HSSP, Evolutionary_Pressure_HSSP, Sting_Relative_Entropy, Sting_Evolutionary_Pressure, Contacts (14 different types of contacts stored in STING_DB and total_energy_of_contacts), Accessible_Surface_in_Isolation and Number_of_Unused_Contact. Using the J48 decision tree algorithm of the data mining software Weka, we searched for those descriptors that define SSE nano-environment. The best results was for 1,867 sequences automatically trained and tested by Cross Validation (10 folds) algorithm for $n=12$. In this case, the Accuracy=81,6%, Precision=54,5%, Sensibility (recall)=43,1% and Specificity=91,1%. We conclude that it is possible to predict the position and content of SSEs based having physicochemical, geometric, structural and conservation descriptors of amino acids.

Word Keys: Prediction, Data Mining, Decision Tree
Supported by: Empresa Brasileira de Pesquisa Agropecuária