Bioinformatics Analysis of Endoglucanase-4 Isolated From Trichoderma Reesei

Ying Wang[1, a], Weihong Yu[1, b] and Dongsheng Yao[1, c]

[1]Jinan University, Guangzhou City, Guangdong Province 510632 China.

[a]yingwang1993@stu2015.jnu.edu.cn, [b]weihong94@foxmail.com, [c]tdsyao@jnu.edu.cn

Abstract

Objective: To predict the basic properties and structure of Endoglucanase-4 using bioinformatics software. Methods: The basic information of Endoglucanase-4 was obtained from the NCBI database; ProtParam, SignaIP 4.1 and NetNGlyc 1.0 Server were used to predict the physicochemical properties, the signal peptide, and potential glycosylation sites. The tertiary structure model of the protein was modeled using Discovery studio and Gromacs. Results: There were 344 amino acids translated from the cDNA of Endoglucanase- 4. The predicted mature peptide was 323 amino acids, its molecular formula was C1480 H2255N399O471S8. And it consists of Gly (10.8%), Thr (10.2%), and Pro (9.9%.). The protein has 20 negatively charged residues (Asp + Glu) and 11 positively charged residues (Arg + Lys). Its molecular weight is 33.43 kDa, its theoretical isoelectric point is 5.12, and its instability index is 28.67, then protein is a stable protein. Its average hydrophilicity index is -0.162, meaning that the predicted protein is hydrophilicity protein. The tertiary structure of Endoglucanase-4 was obtained by homology modeling. Conclusion: The basic structure of Endoglucanase-4 protein was analyzed by bioinformatics, which laid a foundation for the analysis of the structure and function of cellulase family proteins.

Keywords

Endoglucanase-4, Homology modeling, Bioinformatics analysis , Trichoderma reesei.

1. Introduction

Trichoderma reesei can effectively produce cellulases, including endoglucanase, cellobiohydrolase, and glucosidase. Its yield is up to 50% of the total amount of extracellular secretory proteins of Trichoderma reesei. It’s an ideal strain for industrial production of cellulases[1; 2]. Endoglucanase is the most important component of the cellulases. It can hydrolyze soluble cellulose into reducing oligosaccharides[3; 4]. It widely used in foods, feed additives, fabric detergents and enzyme preparations[5-8]. However, the molecular weight, isoelectric point, enzymatic properties, and molecular structure of endoglucanase from different sources and different types also differ. In terms of molecular weight, it is the smallest molecular weight in the cellulases, generally 20 to 50 kDa, some less than 20 kDa; in terms of pH, the pH of most endoglucanase is in the acidic range, which is generally 4 ~5; optimum temperature is generally 50~70°C, and some endoglucanases are 40 °C, for some hydrolyzed cellulose heat-resistant bacteria, the optimum temperature can reach 78 °C. Therefore, the analysis of different sources of endoglucanase is quite critical for its application. This study used bioinformatics methods to predict the structure and properties of Endoglucanase-4, which laid a foundation for further exploration of its biological characteristics.

2. Materials and Methods

2.1 Gene and Protein Sequences

The cel61a gene sequence of Trichoderma reesei QM6a (Accession number: 18188225) and its protein sequence (Accession number: XP_006961567.1) were obtained from NCBI (https: // www. ncbi. nlm. nih. gov/) [9].

2.2 Methods

2.2.1 Physicochemical Properties and Pro (Hydrophobicity) Analysis of Endoglucanase-4

The physicochemical properties of Endoglucanase-4 were analyzed using Protparam (http://expasy.org/tools/protparam.html) in EXPASYP Protemic (http://www.espasy.org). Including the number of amino acids, positive and negative charges residues, molecular mass unit, isoelectric point, molecular formula, stability, etc. Protscale (http://www.espasy.org/cgibin/protscale.pl) was used to analyze the pro (hydrophobicity) of Endoglucanase-4.

2.2.2 Signal Peptide Analysis

Predict the signal peptide sequences using SignalIP 4.1 server (http: // www. cbs. dtu. dk/ services/ SignalIP).

2.2.3 Glycosylation Analysis

The glycosylation site of the Endoglucanase-4 protein was analyzed using NetNGlyc 1.0 Server (http://www.cbs.dtu.dk/services/NetNGlyc/).

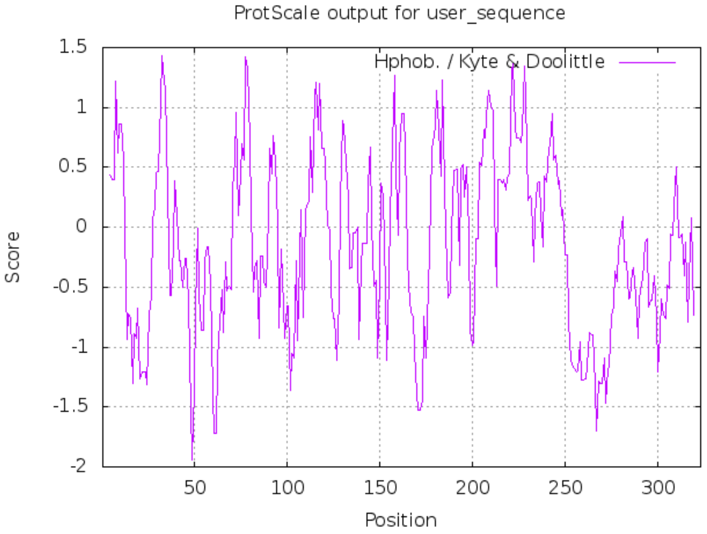
2.2.4 Protein Tertiary Structure Modeling

Established the tertiary structure model of Endoglucanase-4 using Discovery studio 4.5 and Gromacs 4.6.7.

3. Result

3.1 Physicochemical Properties and Pro (Hydrophobicity) Analysis of Endoglucanase-4

There are 344 amino acids translated from the Endoglucanase-4 cDNA, and 323 amino acids are predicted for mature peptides. The molecular formula is C1480H2255N399O471S8. And it consists of Gly(10.8%), Thr (10.2%), and Pro(9.9%.)(Table 1). The protein has 20 negatively charged residues (Asp + Glu) and 11 positively charged residues (Arg + Lys). Its molecular weight is 33.43 kDa, its theoretical isoelectric point is 5.12, and its instability index is 28.67. Then protein is a stable protein. Its average hydrophilicity index is -0.162. And the predicted protein is hydrophilicity protein(Fig.1).



Note: the X-coordinate represents the Amino acid residues, Y-coordinate represents score of hydrophobicity

Fig 1. pro(hydrophobic) water properties of Endoglucanase-4

Table 1. Amino acid composition of Endoglucanase-4

Amino	Number	Ratio(%)
Ala (A)	31	9.60%
Arg (R)	5	1.50%
Asn (N)	21	6.50%
Asp (D)	16	5.00%
Cys (C)	8	2.50%
Gln (Q)	12	3.70%
Glu (E)	4	1.20%
Gly (G)	35	10.80%
His (H)	8	2.50%
Ile (I)	18	5.60%
Leu (L)	19	5.90%
Lys (K)	6	1.90%
Met (M)	0	0.00%
Phe (F)	6	1.90%
Pro (P)	32	9.90%
Ser (S)	27	8.40%
Thr (T)	33	10.20%
Trp (W)	6	1.90%
Tyr (Y)	14	4.30%
Val (V)	22	6.80%
Pyl (O)	0	0.00%
Sec (U)	0	0.00%

3.2 Signal Peptide Analysis

Signal IP4.1server predicts that the protein signal peptide is the first 21 amino acids (Fig. 2), so the mature peptide amino acids are 323.

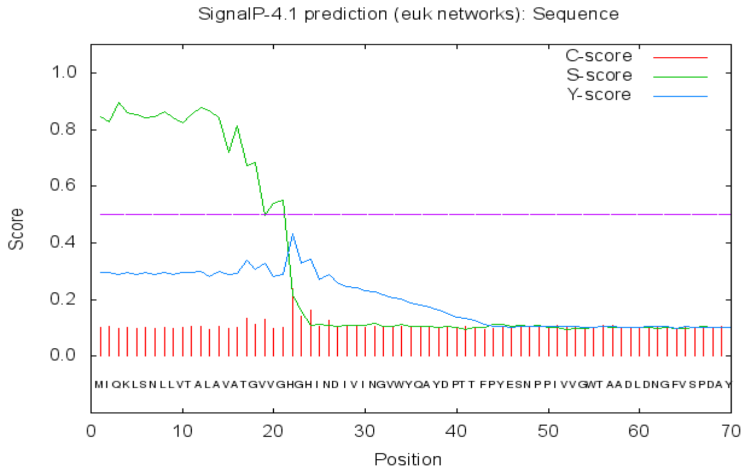


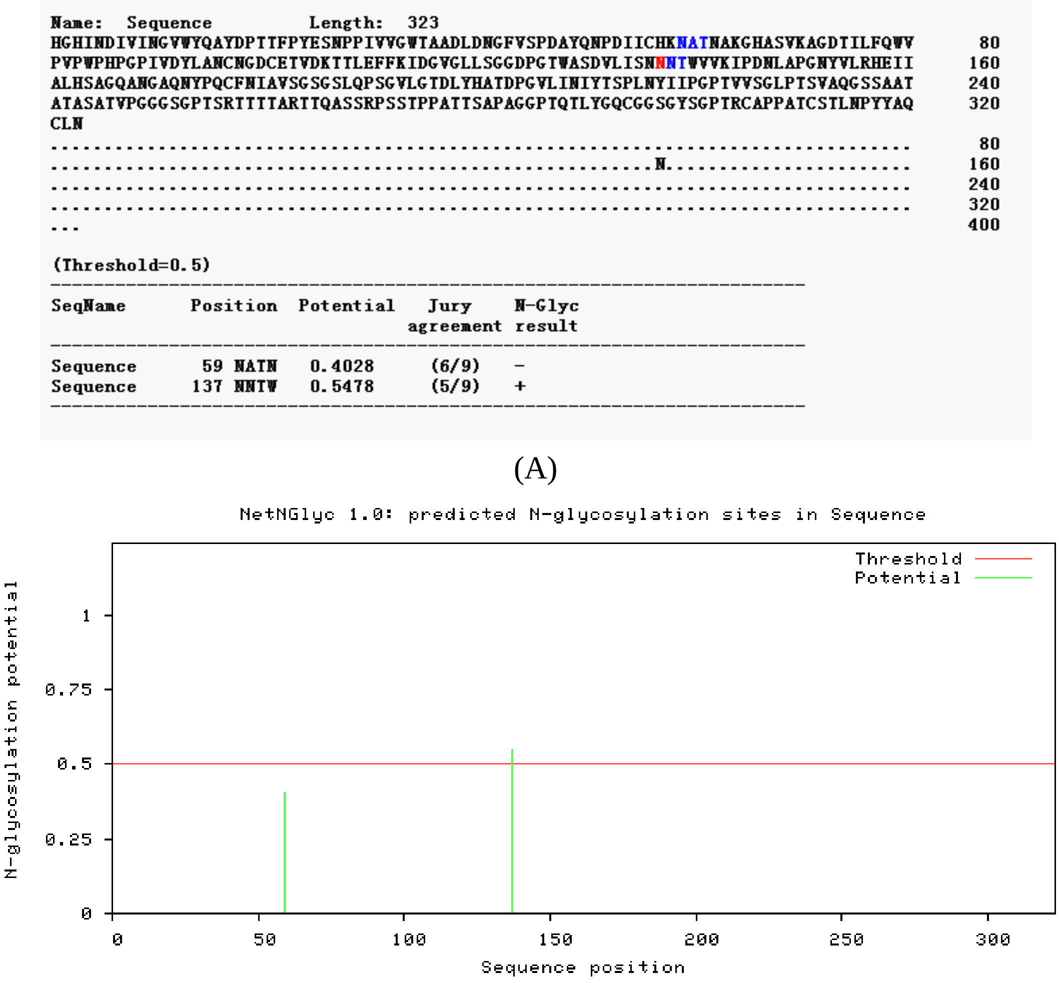
Fig 2. Prediction of Signal peptide of Endoglucanase-4

The mature peptide sequence is as follows:

HGHINDIVINGVWYQAYDPTTFPYESNPPIVVGWTAADLDNGFVSPDAYQNPDIICHKNAT
NAKGHASVKAGDTILFQWVPVPWPHPGPIVDYLANCNGDCETVDKTTLEFFKIDGVGLLS
GGDPGTWASDVLISNNNTWVVKIPDNLAPGNYVLRHEIIALHSAGQANGAQNYPQCFNIA
VSGSGSLQPSGVLGTDLYHATDPGVLINIYTSPLNYIIPGPTVVSGLPTSVAQGSSAATATAS
ATVPGGGSGPTSRTTTTARTTQASSRPSSTPPATTSAPAGGPTQTLYGQCGGSGYSGPTRCA
PPATCSTLNPYYAQCLN

3.3 Glycosylation Analysis

NetNGlyc 1.0 Server predicts that the protein has two potential glycosylation sites, and the 137 amino acid is most likely to undergo glycosylation(Fig.3).



```
Name:  Sequence        Length:  323
HGHINDIVINGVWYQAYDPTTFPYESNPPIVVGWTAADLDNGFVSPDAYQNPDIICHKNATNAKGHASVKAGDTILFQWV      80
PVPWPHPGPIVDYLANCNGDCETVDKTTLEFFKIDGVGLLSGGDPGTWASDVLISNNNTWVVKIPDNLAPGNYVLRHEII     160
ALHSAGQANGAQNYPQCFNIAVSGSGSLQPSGVLGTDLYHATDPGVLINIYTSPLNYIIPGPTVVSGLPTSVAQGSSAAT     240
ATASATVPGGGSGPTSRTTTTARTTQASSRPSSTPPATTSAPAGGPTQTLYGQCGGSGYSGPTRCAPPATCSTLNPYYAQ     320
CLN
................................................................................      80
........................................................N.......................     160
................................................................................     240
................................................................................     320
...                                                                                  400

(Threshold=0.5)
----------------------------------------------------------------------
SeqName      Position  Potential   Jury    N-Glyc
                                 agreement result
----------------------------------------------------------------------
Sequence       59 NATN   0.4028     (6/9)   -
Sequence      137 NNTW   0.5478     (5/9)   +
----------------------------------------------------------------------
```

(A)

NetNGlyc 1.0: predicted N-glycosylation sites in Sequence
Threshold
Potential
N-glycosylation potential
1
0.75
0.5
0.25
0
0
50
100
150
200
250
300
Sequence position

(B)

Fig 3. Prediction of glycosylation sites of Endoglucanase-4

3.4 Protein Tertiary Structure Modeling

3.4.1 Search for Homologous Templates

Using the Protein Model Portal - PSI SBKB (https://www.proteinmodelportal.org/) to search for homology modeling templates (Fig. 4), they are 5o2x, 5o2w, 1azk, and the homology analysis is shown in Table 2.

Fig 4. The homologous templates of Endoglucanase-4

Table 2. Homology analysis of target sequences and templates

Templates	Sequence identity(%)	E-Vale
5o2x	100%	0
5o2w	100%	0
1azk	72%	3.6e^{-8}

3.4.2 Construction of Homologous Models

Homology modeling using Build Homology Models module of Discovery Studio 4.5. The results are shown in Table 3.

Table 3. Homologous model scoring(top ten)

Name	PDF Total Energy	PDF Physical Energy	DOPE Score
M0013	9795.4668	1120.26038	-29016.72266
M0011	9912.3242	1148.760376	-28947.86133
M0005	9925.2012	1131.025741	-28921.93359
M0014	9950.0029	1158.89134	-28866.50977
M0016	9985.9355	1176.700805	-28672.7832
M0009	10010.7637	1175.320344	-29092.66406
M0019	10023.7744	1140.736648	-28775.45117
M0003	10063.7266	1140.207764	-28796.41602
M0010	10082.2285	1160.121622	-28502.20117
M0020	10100.9316	1167.706468	-28663.69141

A total of 20 homology modelling models were built using Modeller. After the model was built, the software automatically scored the energy first.The value of PDF can directly reflect the quality of the constructed model. The smaller the PDF Toal Energey, the greater the credibility of the model. Therefore, we chose the M0013 model for further optimization.

3.5 Model Optimization

Molecular dynamics simulation was used to optimize protein structure, and then to capture the optimal conformation. Finally, the model was evaluated and analyzed. The RMSD values for molecular dynamics simulations are shown in Fig.5. The average conformation after extraction was optimized as shown in Fig. 6.

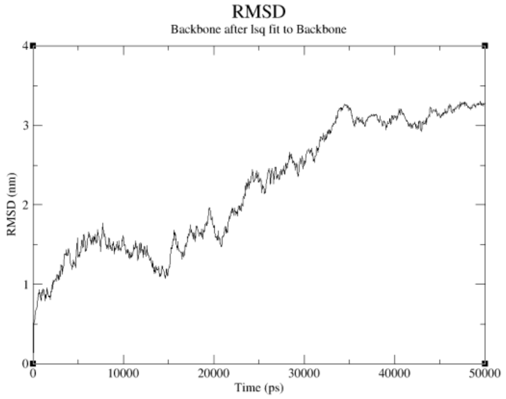


Fig 5. Optimize Endoglucanase-4 molecular structure with MD

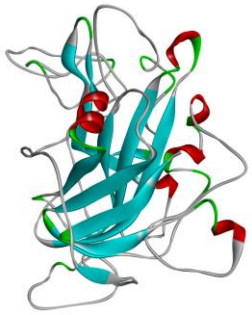


Fig 6. The three-dimensional structure of Endoglucanase-4

3.6 Model Evaluation

Ramachandran's plot suggested that 97.2%, 2%, and 0.8% (Fig.7) of the residues in the derived model were in the acceptable regions, marginal regions, and disallowed regions, respectively. Altogether, 99.2% of the residues were placed into the generously allowed categories, which indicated that the model was reasonable and could be applied for further study.

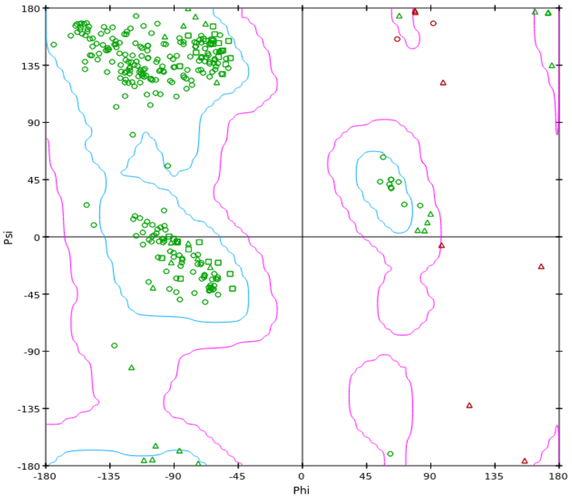


Fig 7. Ramachandran plot of Endoglucanase-4

4. Discussions

Searching homologous templates revealed that the 248-287 amino acids of Endoglucanase-4 haven't homologous templates. Therefore, this study used molecular dynamics simulation to further optimize the constructed model to make the structure closer to the native conformation.

The results of homology modeling showed that the secondary structure of Endoglucanase-4 mainly consists of beta folding, alpha helix, and loop region. Beta folding is the main component, which may be related to the binding of substrate [10-12] .

Prediction of protein structure is one of the hot spots in biological research. It is very important for the study of protein structure and function.

References

[1] Amore A, Giacobbe S, Faraco V. Regulation of Cellulase and Hemicellulase Gene Expression in Fungi[J]. Current Genomics, 2013, 14(4).

[2] Penttilä M, Nevalainen H, Rättö M, et al. A versatile transformation system for the cellulolytic filamentous fungus Trichoderma reesei[J]. Gene, 1987, 61(2): 155-164.

[3] Penttilä M, Lehtovaara P, Nevalainen H, et al. Homology between cellulase genes of Trichoderma reesei: complete nucleotide sequence of the endoglucanase I gene[J]. Gene, 1986, 45(3): 253-263.

[4] Juturu V, Wu J C. Microbial xylanases: engineering, production and industrial applications[J]. Biotechnology Advances, 2012, 30(6): 1219-1227.

[5] Karmakar M, Ray R R. Current Trends in Research and Application of Microbial Cellulases[J]. Research Journal of Microbiology, 1994, 6(1).

[6] Narra M, Dixit G, Divecha J, et al. Production, purification and characterization of a novel GH 12 family endoglucanase from Aspergillus terreus and its application in enzymatic degradation of delignified rice straw[J]. International Biodeterioration & Biodegradation, 2014, 88(4): 150-161.

[7] Bernardi A V, De Gouvêa P F, Gerolamo L E, et al. Functional characterization of GH7 endo-1,4-β-glucanase from Aspergillus fumigatus and its potential industrial application[J]. Protein Expression & Purification, 2018.

[8] Kuhad R C, Gupta R, Singh A. Microbial Cellulases and Their Industrial Applications[J]. Enzyme Research,2011,(2011-9-7), 2011(2): 280696.

[9] Martinez D, Berka R M, Henrissat B, et al. Genome Sequencing and Analysis of the Biomass-Degrading Fungus Trichoderma reesei (syn. Hypocrea jecorina)[J]. Nature Biotechnology, 2008, 26(5): 553.

[10] Karlsson J, Saloheimo M, Siika-Aho M, et al. Homologous expression and characterization of Cel61A (EG IV) of Trichoderma reesei[J]. Febs Journal, 2010, 268(24): 6498-6507.

[11] Pierce B C, Agger J W, Wichmann J, et al. Oxidative cleavage and hydrolytic boosting of cellulose in soybean spent flakes by Trichoderma reesei Cel61A lytic polysaccharide monooxygenase[J]. Enzyme & Microbial Technology, 2017, 98: 58-66.

[12] Tanghe M, Danneels B, Camattari A, et al. Recombinant Expression of Trichoderma reesei Cel61A in Pichia pastoris : Optimizing Yield and N-terminal Processing[J]. Molecular Biotechnology, 2015, 57(11-12): 1010-1017.